Advanced statistical methods for genetic studies of lupus

James Roberts

A thesis presented for the degree of
Masters of Science in Statistics

Department of School of Mathematics
Statistics and Actuarial Science
University of Kent
England
2020

Acknowledgements

Dedicated to my academic superstar niece, Georgia, whose university time will come later in life than originally expected, just like mine. You will beat the battle of chronic illness and go on to a successful career just as successful as your time was at school.

"...and here you are, continuing on despite how hard it's been"

The greatest of thanks to my supervisor Dr James Bentham for guiding, educating, supporting, and putting up with me during the unprecedented year of 2020.

Also, a big thank you to Claire Carter. Always available for a chat or advice when needed. If it wasn't for Claire, I wouldn't have had my study time at the University of Kent.

Another big thank you goes to Derek Baldwin for all his technical support of changing my operating system and giving me the ability to work remotely.

Thank you to all at my Statistics reading group, including Sara, Mariza, Peter and Ulrike.

For moral support during such a year of global disarray, thanks go to my family and friends especially Andreea, Bill, "Daft" and Danielle.

...and finally my dearest thanks go to my wife Kellie and my son Ryan for making life so much easier after re-entering education.

Love to you all

Abstract

Systemic Lupus Erythematosus (SLE) is a complex autoimmune disease that has a large genetic component. It has been researched multiple times using Genome Wide Association Studies (GWAS), mainly with European and Asian cases. These studies have tested and analysed genetic markers separately and are simplistic statistically, with strict corrections made for multiple testing to reduce the false positives that they produce. In 2015, the largest GWAS of SLE at that time was investigated by Bentham et al. Using the data employed in their research, the goal of the work presented here was to complement the study's results by applying sophisticated statistical methods to the data whilst reducing the number of false positives. Four methods were implemented with the data, two Bayesian approaches and two frequentist. Three methods implemented a mixture of optimization and regularization techniques, whilst the other uses a standard frequentist association test. The methods were sporadic in finding the published associated hits and lacked consistency. Dense blocks of SNPs that were in high linkage disequilibrium were reduced to selecting just one or a few SNPs to represent the associated risk allele, which is a clear advantage of these methods. Although the results were inconclusive, it was noticeable, that on average, the percentage of non-zero coefficients chosen by the variable selection methods grows as the chromosomes get smaller in size which is counterintuitive, and may be an undesirable artefact produced by these methods. Furthermore, on three occasions (chromosome 3, 19 and 22), methods have chosen numerous non-zero coefficients per chromosome in comparison to the other methods. Overall, further research is required, to produce a consistent and reliable model using variable selection techniques, that reduce false positives and reveal novel associated hits that ultimately result in discovery of casual SNPs in the fight against disease. This work is a step towards this goal.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Introduction

With the advent of technological advancements from the early microarrays to the invention of next generation DNA sequencing technology, we now have a far better understanding of the workings of the human genome. This has enabled us partially to grasp the complexities of the building blocks of life. The human genome holds all the information for us to function and live, with around 99.9% of DNA the same for every person [1]. A variation in the genome affects an individual's risk of disease and response to medicine [2]. Since the mid 1990s the evolution in superior techniques has delivered on finding rarer variants that are associated with causes of diseases through improved coverage of the genome. The revolution of personalised medicine has begun, with whole genomes being sequenced in just a day with the cost decreasing rapidly and so the number of datasets will continue to grow [3]. Through increased work in mathematics, statistics, and computer science a new era has begun in the form of bioinformatics and biostatistics.

This thesis presents biostatistics analysis based on data for systemic lupus erythematosus (SLE) from Bentham et al (2015) entitled "Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus" [4]. The overall aim of this work is to assess whether advanced statistical models can be applied to a practical dataset to find associations with the disease. The main objective of the study is to assess the chosen statistical methods. In the first instance, this is to assess whether the chosen statistical methods perform well in finding SNPs or loci that are already known to be associated with the disease. Secondly, it is necessary to assess whether the methods produce large numbers of false positive results. The long term goal is to replace the existing simplistic statistical models used by large genetic studies, but this is beyond the scope of this thesis as this would be a major research project. The overall literature gap is between the large-scale genetic studies that use relatively simplistic statistical models, and sophisticated statistical models that have not been tested on large, practical genetic datasets.

1.1.1 Genetic Disease and Personalized Medicine

Diseases that are classified as genetic are caused by the variations in DNA sequences [5]. Some variants increase the risk while others decrease the risk of disease but most cause no effect at all. These changes are called genetic variants. Some genetic variants can lead to single gene disease like cystic fibrosis while others like lupus are a complex disease caused by multiple variants interacting with environmental influences [6]. Studies in genetics are providing a foundation for future work into the eventual goal of personalized medicine. This would help to target the correct drugs to administer to the patient, can help in the diagnosis of a disease at an earlier stage than was previously possible, and help with the prediction of risk of contracting a variety of diseases [1]. Due to

the uniqueness of a person's DNA these will become standard practice in the future, reducing costs of treating long-term illness by optimization of therapies resulting in the individual health needs of patients around the world becoming easier and cheaper to diagnose whilst saving millions of lives. This is the motivation for work in this field.

1.1.2 Original data set

In 2015 Bentham et al [4] set out to unearth associations of novel susceptibility loci with the autoimmune disease systemic lupus erythematosus (SLE) based on European ancestry. The study comprised of a fresh GWAS of 4946 cases (before data cleaning) with SLE and 1286 controls that were added to 5727 other healthy controls taken from the University of Michigan Health and Retirement Study. Additional data from Hom et al [7] was utilized, containing an extra 1165 cases and 2107 controls. Imputation software Impute [8] and SHAPEIT [9] was used on both sets of data to the density of the 1000 Genomes project [10], however, only the genotyped data from the GWAS was analysed in this work. The paper was the largest GWAS on lupus at the time. The statistical modelling in this study involved single SNP logistic regression that did not take in account any linkage disequilibrium. In statistical terms, this analysis was therefore somewhat simplistic. Each of the cleaning steps for GWA studies described in section 1.3 were carried out on this dataset, and so the data used in this thesis did not require any additional cleaning, i.e., the dataset is that used in the Bentham et al [4] study without any modifications, with the cleaning steps having been considered by the reviewers of that paper.

1.1.3 Study Objectives

Using the 4036 SLE cases and 6959 controls with data comprising of 644,674 SNPs of post-cleaned data from the original research by Bentham et al [4], this study will attempt to go further statistically using advanced variable selection techniques, with the aim to produce a relatively small number of non-zero coefficients to analyse that also accounts for linkage disequilibrium and potentially recording fewer false positives. This thesis will aim to compare and assess the performance of the statistical models chosen, that can be capable of processing this large data set by different frequentist and Bayesian methods. This thesis can also be used as a replication study in conjunction with any potential findings of associations that have not been shown to be prevalent in European populations but are known in Asian and African populations. Comparisons with the previously known associated SNPs will be made in the results section which will be insightful to whether the methods have successfully found the hits, producing confidence in the results. The standard techniques of GWAS in the study of SLE use PLINK [11] and SNPTest [12] and have resulted in successful associations. It is estimated that over 50% of the disease's heritability can be explained from the 13 GWAS studies of European and Asian ancestry, resulting in 84 loci that

have been associated with the disease [13]. The future requires new methods to gain further advancement into the missing heritability that remains. This could be explained by rare variants that fail to produce low enough p-values to be classed as statistically significant with the standard measure of genome wide significance of p-values at $p < 5 \times 10^{-08}$. There are also linkage disequilibrium (correlation) problems that can potentially be better understood by using variable selection methods. This thesis attempts to trial advanced statistical methods with the potential that one may become a new standard.

1.1.4 Statistical methods of previous lupus GWAS

Genome-wide association studies involve hundreds of thousands of different SNPs to analyse, and need to take into account linkage disequilibrium amongst those SNPs, therefore GWA studies are complicated statistical problems that have no ideal format after years of studies performing many types of methods [14]. Very little literature has appeared with regularization methods that have studied the disease lupus in the pursuit of finding phenotypic variation of complex traits. Unfortunately the types of GWAS carried out to date will not point out the rare variants that lie beneth the genetic p-value threshold of 5×10^{-08}, although a paper in 2019 reported 24 previous GWAS on Lupus and a total of 388 suggestive associations made that were below the genetic threshold [15]. This thesis will try to fill some of the current void.

1.2 Genetics

1.2.1 Introduction

The study of genes and heredity was formed in the 1800s by an Augustinian monk, Gregor Mendel [16], whose work into pea plants and their inheritance structure began the basis of what we know as genetics. The expansion of the field over the past 150 years has revolutionised the way we think about every living cell on Earth. A paper from 2000 estimates that a human has around 28,000 to 34,000 genes [17]. It is thought that not all these are protein coding genes. Genes that do not code for proteins are called functional RNA. The whole set of genes in a cell is called the cell's genome. Nucleotides are made up of a phosphate group, a pentose sugar, and a nucleobase, and it is these nucleotides that make up DNA and RNA. DNA is used to store genetic information and their molecules are extremely long and are compacted into a cell's nucleus [18]. Each DNA nucleotide has one of a possible four bases - A Adenine, T Thymine, C Cytosine and G Guanine. A gene is a sequence of DNA bases. RNA transfers genetic information from the DNA to the ribosomes [5].

1.2.2 Chromosomes

There are 23 pairs of chromosomes in most humans, 22 autosomal chromosomes (1-22 in order of size with 1 being the largest) and a pair of sex chromosomes (generally XX for female and XY for male). The Y chromosome is smaller than

the X chromosome and thus carries fewer genes. Each chromosome is made up of DNA wrapped around proteins called histones (as shown in Figure 1). There are two sections in a chromosome divided by the centromere. One arm has a side named q (the longer of the two) and the other is called p. They have hundreds and occasionally thousands of genes [5].

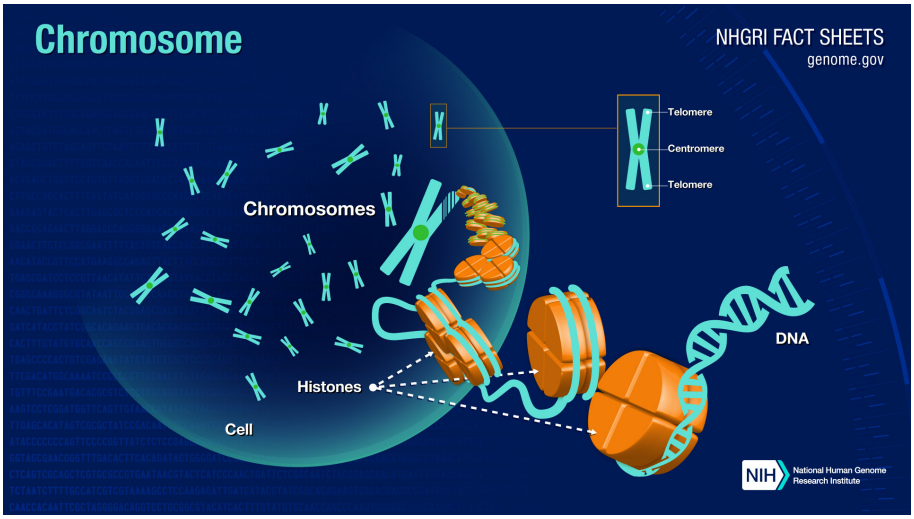

Figure 1: Chromosome Facts: Source - National Human Genome Research Institute [1]

The chromosomes 1-22 are almost the same size in each person and have the same genes, but there can be different versions of a single gene, called alleles. Each human has the same genes although they have different alleles. The alleles in a gene are what brings about genetic diversity. Genes are varied in size and can be coding or non-coding for proteins. Genes are inherited one from each biological parent [5]. Deletions, Duplications, Translocations, Inversions and Rings are all alterations that can happen to a chromosome's structure [1] as shown in Figure 2.

1.2.3 Single Nucleotide Polymorphisms

A single nucleotide polymorphism (SNP) is a point mutation that usually takes place during DNA replication that produces a single base pair. This occurs when the DNA sequence differs from the majority of the population. There are around 4-5 million SNPs in a human genome [19], and it is estimated that they account for 90% of genetic variation in the genome. The high frequency of SNPs occurring makes it possible for high density profiling to be tackled. Each position of a SNP has its own identity and is encoded by a marker beginning with rs.

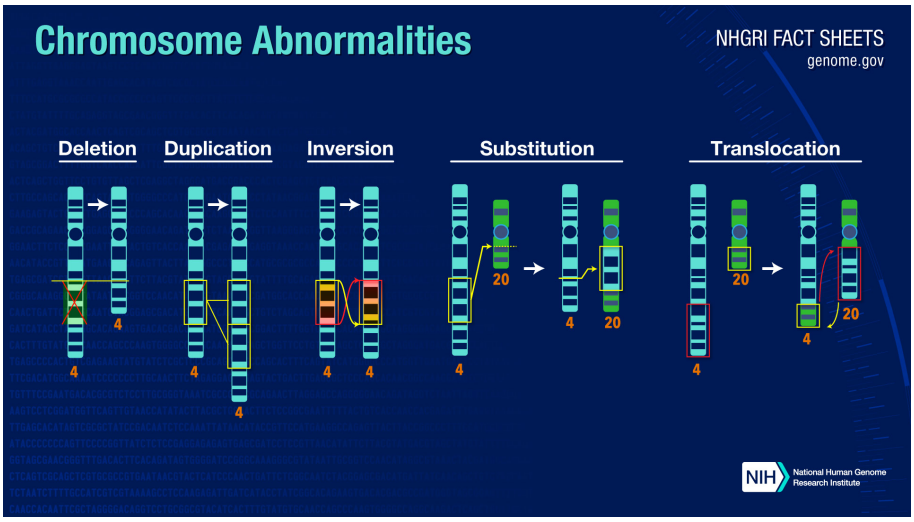

Figure 2: Chromosome Abnormalities: Source - National Human Genome Research Institute [1]

SNPs can be positioned in:

- Coding sequence of a gene
- Non-coding regions of a gene
- Intergenic regions between genes.

1.3 Genome Wide Association Studies

1.3.1 Introduction

GWAS has been used as a statistical analysis tool for 18 years, with its origins in a 2002 paper researching myocardial infarction [20]. This was followed up in 2005, when a study was produced for age related macular degeneration by Klein et al [21]. Since then and up to 29th October 2018 there have been 3639 research papers that have been published across 3508 unique diseases tested through the GWAS format with thousands of associations with disease having been made [22]. GWAS are designed to detect and analyse individual or regional areas of SNPs that are potentially associated with disease. Although the causality cannot be defined by these types of studies, they are beneficial for helping other research (often laboratory based) to determine the true causality associated with the disease in question. Rarer genetic variants with small effects are being searched for, through regional association analysis rather than individual SNP analysis [23]. This new method through exome and whole genome sequencing of rare variants has the potential to discover causation of complex diseases.

1.3.2 Allele Frequency

At each locus, two (and occasionally three) alleles are observed. The second most common allele that is observed at that particular locus is denoted as the minor allele and it's count can be divided by the total amount of alleles to produce a minor allele frequency as shown in Equation (1).

$$\text{Allele frequency} = \frac{\text{number of copies of the allele}}{\text{number of all copies of the allele at the locus}} \tag{1}$$

SNPs are of huge significance in genetics, particularly in comparing cases and controls in GWA studies. Early GWA studies were based around SNPs with minor allele frequency (MAF) of less than 5%. As time has moved on, the studies now use a MAF of less than 1%. As databases increase in size, so will population study numbers and the frequency thresholds will continue to drop.

1.3.3 Population Stratification and Cryptic Relatedness

When performing Genome Wide Association Studies, a systematic difference in allele frequencies between subpopulations arises. This is called population stratification. The two main types of statistical GWAS are family-based design and unrelated population-based design. It is in the latter in which population stratification can lead to false positive and false negative findings if not accounted for. Although it is easier to collect samples on unrelated individuals rather than relations in a family, there have been cases where studies have produced spurious associations through variations in ancestry. There are 3 main types of genetic population structures globally: European, Asian, and African. There is also much diversity amongst minor subpopulations too. The mixing of people through migration has resulted in complications for genetic based studies. For example, the genetic diversity in Brazil is very high. This has formed over time by migration from Europeans, native Americans and Africans resulting in a high multiracial mix. The GWAS landscape has been dominated by studies of European ancestry with 86.0% discovery and 76.7% of replications with most of the remainder being from Asian subjects with the majority of funding for GWAS coming through the United States of America (85.1%) and the United Kingdom (14.4%) [22].

Genetic diversity between populations in case-control studies is a problem due to SNP alleles that are common in one geographical or ethnic group being much rarer in another. Small differences in the effect of alleles may produce confounding.

A strict part of a GWAS quality control now must contain standard population stratification methods. To deal with population stratification, cryptic relatedness and missing genotype rates are dealt with by quality control using software such as EIGENSTRAT [24] or PLINK [11].

When dealing with population stratification in case control studies, cryptic relatedness needs to be a consideration. Close relatives that are not known to the study organisers could well confound the results. A study in 1999 by

Devlin and Roeder [25] reported that cryptic relatedness was a source of confounding in studies. The independence of populations is brought into question if close relatives are found in the data and thus the design of the study must be adjusted. Although using family-based data is used less in design methods, it has more control over population stratification than a design based on population studies but is simply not practical if sample sizes reach 1000 or more.

1.3.4 Cleaning Data and Imputation

Cleaning of the data is a key part of a successful GWAS. Quality control steps are taken to remove markers with high error rates. There will be expected to be numerous imperfections including poor DNA hybridization to the array, some samples of DNA not being of high enough standard and there may even be contaminated samples.
Checking the consistency of sex, minor allele frequency exclusions, sample relatedness, population stratification, heterozygosity rates, missingness of SNPs and Hardy Weinberg equilibrium deviation rates are all analysed in carrying out quality control of the uncleaned data. Software packages like PLINK [11] and SNPTest [12] clean the data under supervision of the researcher. This dramatically reduces the number of SNPs with significant associations that studies must further analyse. Imputation is required when there are missing data (for example when datasets use different genotyping chips) or unobserved genotypes in a GWA study. Imputation can improve the statistical power in a single study or meta-analysis where causal SNPs are not present on the genotyping chip but are in the 1000 Genomes dataset, for example. The method is based around known levels of linkage disequilibrium (i.e. correlation) using knowledge of haplotype structure. Imputation software like Impute [8] and SHAPEIT [9] link up with reference panels from HapMap [26] or the 1000 Genome project [27]. However, in the case of the data analysed in this thesis, only directly genotyped data was used and the dataset had been cleaned previously, so no further cleaning was carried out by the author. There were no SNPs with LD=1 in the final cleaned dataset and so there were no need to allow for this in these analyses. LD is taken into account implicitly by the regression models. There is no missingness in the data, and the methods were each applied to the same dataset.

1.3.5 Hardy Weinberg Principle

The Hardy-Weinberg principle states "that under the condition of large population size, diploid organisms with non-overlapping generations and random mating, the genotype frequencies at a locus are determined by the allele frequencies, and both the genotype and the allele frequencies will stay constant in future generations when the conditions of no mutation, no migration and no selection hold" [28].

The Hardy-Weinberg Equilibrium test (HWE) is a vital part of statistical analysis in population studies involving genetics based on the Hardy-Weinberg

Principle. It assesses the allele frequency changes through generations. These can be caused by mutations, deletions, genetic drift, inbreeding and cryptic relatedness. The allele frequencies should be passed down with the same ratio. Based on a single gene with only two alleles, with the minor allele represented by a, and the major allele represented by A, Equation 2, presents the expected genotype frequencies under random mating for AA, Equation 3 for aa and Equation 4 for Aa, with the frequency of observed alleles are represented by p and q, where $q = (1 - p)$ and where $p^2 + q^2 + 2pq = 1$.

$$f(AA) = p^2 \tag{2}$$

$$f(aa) = q^2 \tag{3}$$

$$f(Aa) = 2pq \tag{4}$$

$$p, q \in [0, 1]$$

also as shown in Table 1 (which is known as a punnet square) with the allele from both parents:

Table 1: Hardy-Weinberg Punnet Square

Calculations of expected allele frequency		
	PATERNAL	PATERNAL
MATERNAL	$AA(p^2)$	$Aa(pq)$
MATERNAL	$Aa(pq)$	$aa(q^2)$

From the punnet square we record the expected allele frequencies (from the control group) and compare them to the observed allele frequencies (from the case group). If the values are considerably different then it is said to not be in HWE. This test was carried out in the original study as part of the data quality checks.

1.3.6 Cochran Armitage Test

In case-control studies for complex traits the standard test of genetic association is the Cochran Armitage trend test [29, 30]. This standard procedure for single SNP analysis uses a null hypothesis of no association between a disease risk and a genotype. Different genetic models can be used in case control analysis including Full genotype model, Dominant model, Recessive model, Multiplicative model and finally the Additive model [31]. In genetic studies the additive model is considered the most realistic representation of genetic risk and so is preferred and thus is used in this thesis.

An additive model consists of homozygous genotype coded 0 (0+0=0), heterozygous 1 (0+1=1, or 1+0=1) and another homozygous genotype 2 (1+1=2) and it is thought that in an additive model the risk changes when there is a potential exposure of risk.

With the use of contingency tables the additive model is used to calculate the risk of disease based on the amount of risk alleles per loci. For risk allele (A) and Non risk allele (a): With homozygous $aa = (0)$ there are no risk alleles (weights of 0), heterozygous $aA = (1)$ one risk allele and homozygous $AA = (2)$ two risk alleles. It is assumed in this model that Aa has a risk factor of r-fold for the risk of developing the disease and AA is 2r-fold (double the risk factor) of the risk. The test results in a set of scores and can be used as a score test. All four methods used additive methods in this study.

1.3.7 Haplotypes and Linkage Disequilibrium

Haplotypes are a group of SNPs that are inherited together in blocks across a chromosome. They are passed down through generations by genetic recombination, although this does not occur so often inside each block because they are positioned very close to each other. The haplotypes that appear in blocks should have higher levels of linkage disequilibrium (i.e. correlation), and this can help or hinder any GWA study. The alleles can be inside different genes or even between genes. GWAS will highlight SNPs with true genetic associations. With the use of software like Plink [11], block analysis and linkage disequilibrium can be calculated.

Linkage is when two alleles are located on the same chromosome therefore, they are physically linked together. If they are in linkage equilibrium then the observed frequencies will be the same as the expected frequencies, thus there is an equal probability of inheriting each allele. For linkage disequilibrium it can be assumed that the observed frequencies (f) are different from the expected frequencies.
The linkage disequilibrium coefficient is D' and is calculated as in Equation 5.

$$D' = \frac{D}{|D|_{max}} \tag{5}$$

where $D = f_{AB} \times f_{aB} \times f_{Ab} \times f_{ab}$ and $D_{max} = min(f_A \times f_b, f_a \times f_B)$.

If $D > 0$, $|D|_{max}$ is equal to the smaller value $f_A \times f_b$ and $f_a \times f_B$.
If $D < 0$, $|D|_{max}$ is equal to the smaller value $f_A \times f_B$ and $f_a \times f_b$.

r^2 is used for the correlation between loci with $r^2 = 0$ (linkage equilibrium) and $r^2 = 1$ (complete linkage disequilibrium).

1.3.8 GWAS Limitations

It is perceived that there are multiple rare or low frequency variants that cannot be identified by a GWAS owing to a lack of statistical power. It is these rare multiple traits analysed together that have small effect sizes [32]. It was reported by Yang et al [33] that many SNPs gave such small effects that they were not significant enough to be noted in a GWAS and it was this that was causing the missing heritability proposed by Maher [34]. Yang et al also reported that some

or all of the genetic mutations are not in perfect linkage disequilibrium and this was another possible reason for missing heritability. Yang et al results noted causal variants have lower MAF than common SNPs and their frequencies are too small to be picked up by a GWAS. Some scientists believe the focus should be on ultra-rare variants and researched through other techniques like differential network analysis using next generation sequencing [35]. Another limitation is the standard association testing that has been used frequently in large data environments. More sophisticated techniques can be used to reduce processing times and the vast amount of false positives that these tests produce. This is a key motivation for the work presented in this thesis.

1.4 Genetics of Systemic Lupus Erythematosus

1.4.1 Introduction

Systemic lupus erythematosus (SLE) is a complex autoimmune disease with strong genetic and environmental makeup. There are four types of lupus: Systemic lupus erythematosus, the most common form, Cutaneous lupus erythematosus, Drug-induced lupus erythematosus and Neonatal lupus [36]. It is a case of an autoimmune disease that attacks healthy cells inside the body using its own immune system. It has been assumed that SLE is ignited by environmental factors like smoking, UV light and alcohol consumption [37]. Hormonal factors in females that are pregnant or are on hormone replacement therapy can bring on SLE in genetically susceptible people. It has a range of severity from mild cases like fever, joint pain, and rash to potentially life threatening conditions affecting major organs like kidney disorder, heart attacks and strokes if not diagnosed early [36]. Research estimated around 12 percent of people with lupus will prematurely die from lupus complications [38]. Estimations from the Lupus Foundation of America notes that a possible 5 million people worldwide have lupus [39]. It is prone to be in females aged 15-44 (child-bearing age), particularly in women that are of an African, Asian, or Caribbean origin. Lupus is primarily a female disorder but can also affect males too with a ratio of 9:1. This shows that gender and population group are a major factor when considering any study on lupus [40].

1.4.2 GWAS in SLE

The first two main genome wide association studies involving SLE were by Harley et al [41] and Hom et al [7] both in 2008. Harley et al researched 720 females for the study with 2337 controls all of European ancestry. All cases were considered under classification of SLE from the American College of Rheumatology and stated the diagnosis of SLE is 4 or more of their criteria must be met [42]. This is presented in Table 2.

Hom et al [7] studied 1435 of both female and male cases with 3583 controls all from a European descent with replication through Swedish case-controls. Both studies found association with replication: Hom et al found SNP rs13277113

Table 2: SLE diagnosis criteria from the American College of Rheumatology.

4 or more of the following criteria:
Anemia
Anti-nuclear antibodies
Arthritis
Cardio Pulmonory involvement
Immunological disorder
Kidney disorder
Malar Rash
Mouth or nose ulcers
Neurological disorder
Photosensitivity
Skin rash.

between the genes *BLK* and C8orf13 in chromosome 8 and between *ITGAM* and *ITGAX* on chromosome 16 in SNP rs11574637. Harley et al [41] found 4 regions that were associated with SLE: On chromosome 16 *ITGAM*, on chromosome 11 *KIAA1542*, chromosome 3 *PXT* and the SNP rs10798269 on chromosome 1. The majority of loci that have been associated with SLE are in European and Asian population studies. In 2016, a meta-analysis of these studies showed that over half of the published SLE genetic associations are present in both populations [43]. SLE has been associated with many reported loci but despite this, a large slice of genetic component remains to be discovered [13].

The standard statistical methods used in GWAS were also used in the studies involving SLE, and all used association testing based on logistic regressions, with disease status being predicted by genotype. They applied additive single SNP models, using PLINK [11] and SNPTest [12] for association testing. Bentham et al pooled the datasets of Hom et al, Harley et al and their own study to form a meta-analysis which increased the power of the study but for each of these studies used single SNP logistic regression models, which is still the standard for GWAS.

1.4.3 Major Histocompatibility Complex

The major histocompatibility complex (MHC) lies within 6p22.1 to 6p21.3 on the short arm in chromosome 6 and contains 224 genes spanning 3.6 mega base pairs [44]. First associations of a genetic link to SLE were reported in the MHC in 1971 [45, 46]. The MHC region of genes plays a crucial role in susceptibility to autoimmune diseases and studies for lupus have previously concentrated in this area [2, 47, 48], see Figure 3. The MHC is the densest part of the human genome. This region of genes has high linkage disequilibrium amongst SNPs, and this causes problems in identifying single SNPs for association [49].

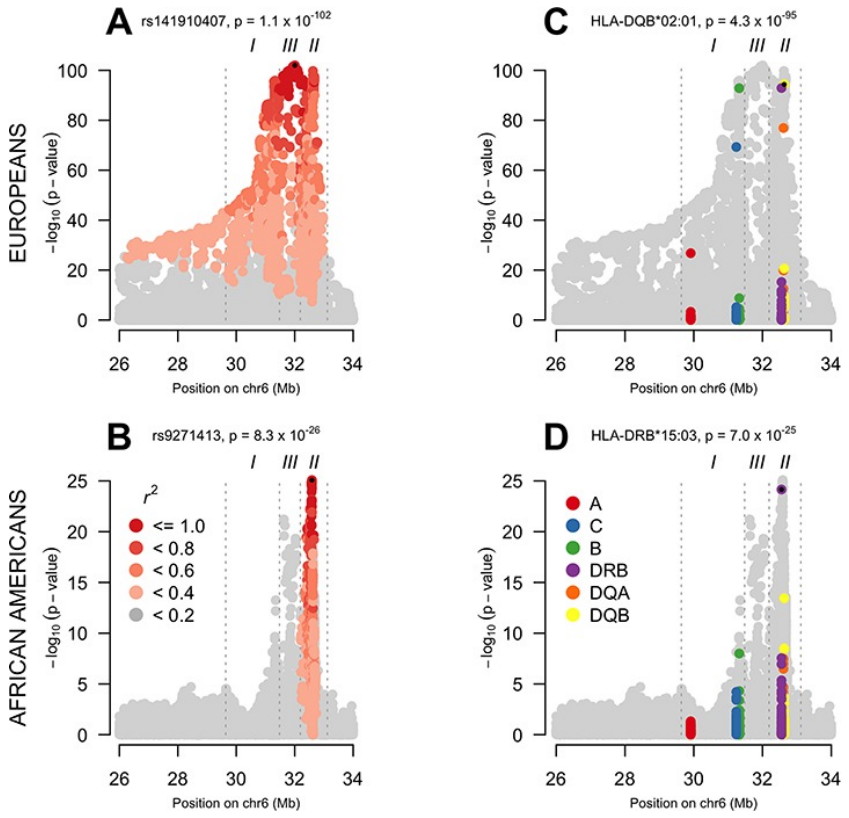

Figure 3: Manhattan plots showing a thick correlated region of SNPs that make up the complex region MHC partly using Bentham et al European data and Hansombe et al African American data. The title of each panel presents the SNP with the lowest p-value score with correlation of the other markers shown by the r^2 values. The darkest shaded dot is the most significant marker with LD shown using pink and red. The coloured dots represent genes in genes within the MHC, with alleles for DRB denoted by asterisks and MHC classes denoted by I, II and III. Source: Hanscombe et al [2]

1.4.4 Associated Genes to SLE in European populations

Genes that were associated with SLE pre-GWAS were: *PTPN22* (chr 1) [50], *FCGR3A* (chr 1) [51], *FCGR2A* (chr 1) [51], *STAT4* (chr 2) [52], *PDCD1* (chr 2) [53], *TREX1* (chr 3) [54], *SPP1* (chr 4) [55], *BANK1* (chr 4) [56], *HLA-DRB1* in the HLA region (chr 6) [57], *IRF5* (chr 7) [58]. SNPs associated with lupus through GWA studies are shown in Table 3 and Table 4.

Table 3: Timeline of asscociated SNPs with lupus through GWAS 2008-2014.

GWAS 2008-2014					
Year	Chr	Associated SNP	Likely causal gene	Study population	Author
2008	1	rs10798269	*TNFSF4*	EUR	HAR
2008	3	rs6445975	*PXK*	EUR	HAR
2008	6	rs5029939	*TNFAIP3*	EUR	GRA
2008	8	rs13277113	*BLK,C8orf13*	EA	HOM
2008	11	rs4963128	*KIAA1542*	EUR	HAR
2008	16	rs11574637	*ITGAX*	EA	HOM
2008	16	rs9888739	*ITGAM*	EUR	HAR
2009	1	rs3024505	*IL10*	EA	GAT
2009	5	rs7708392	*TNIP1*	EA	GAT
2009	6	rs6568431	*PRDM1*	EA	GAT
2009	6	rs11755393	*UHRF1BP1*	EA	GAT
2009	7	rs849142	*JAZF1*	EA	GAT
2011	1	rs525410	*LAMC2*	EUR	CHU
2011	4	rs4956211	*COL25A1*	EUR	CHU
2011	6	rs1150754	*TNXB*	EUR	CHU
2011	5	rs2431697	*PTTG1*	EUR	CHU
2014	1	rs10911628	*EDEM3*	EUR	ARM
2014	2	rs12993006	*BIN1*	EUR	ARM
2014	2	rs4544377	*KCNJ3*	EUR	ARM
2014	3	rs4684256	*CNTN6*	EUR	ARM
2014	7	rs6946131	*SEC61G*	EUR	ARM
2014	10	rs10857712	*MTG1*	EUR	ARM
2014	11	rs10466455	*EHF*	EUR	ARM
2014	15	rs11073328	*FAM98B*	EUR	ARM
2014	15	rs12259	*TYRO3*	EUR	ARM
2014	15	rs8023715	*SPATA8*	EUR	ARM
2014	17	rs11655550	*MED1*	EUR	ARM
2014	20	rs6084875	*RASSF2*	EUR	ARM
2014	20	rs11697848	*RNF114*	EUR	ARM

KEY: EUR=European, EC=European and Chinese, EA=European American, HAR=Harley et al [41], GRA=Graham et al [59], HOM=Hom et al [7], GAT=Gateva et al [60], CHU=Chung et al [61], ARM=Armstrong et al [62].

Table 4: Timeline of asscociated SNPs with lupus through GWAS 2015-2020.

GWAS 2015-2020					
Year	Chr	Associated SNP	Likely causal gene	Study Population	Author
2015	2	rs67040462	*SPRED2*	EUR	BEN
2015	2	rs3768792	*IKZF2*	EUR	BEN
2015	3	rs564799	*IL12A*	EUR	BEN
2015	5	rs7726414	*TCF7,SKP1*	EUR	BEN
2015	11	rs3794060	*DHCR7,NADSYN1*	EUR	BEN
2015	12	rs10774625	*SH2B3*	EUR	BEN
2015	14	rs4902562	*RAD51B*	EUR	BEN
2015	16	rs9652601	*CIITA,SOCS1*	EUR	BEN
2015	17	rs2286672	*PLD2*	EUR	BEN
2015	X	rs887369	*CXorf21*	EUR	BEN
2016	1	rs34889541	*PTPRC(CD45)*	EC	MOR
2016	1	rs2297550	*IKBKE*	EC	MOR
2016	2	rs7579944	*LBH*	EC	MOR
2016	2	rs17321999	*LBH*	EC	MOR
2016	3	rs6762714	*LPP,TPRG1-AS1*	EC	MOR
2016	6	rs17603856	*ATXN1*	EC	MOR
2016	6	rs597325	*BACH2*	EC	MOR
2016	7	rs73135369	*GTF2IRD1-GTF2I*	EC	MOR
2016	9	rs1887428	*JAK2*	EC	MOR
2016	11	rs494003	*RNASEH2C*	EC	MOR
2016	16	rs1170426	*ZFP90*	EC	MOR
2017	4	rs3733345	*DGKQ*	EA	LAN
2017	6	rs10498722	*LRRC16A*	EA	LAN
2017	6	rs4712969	*SLC17A4*	EA	LAN
2017	6	rs2327832	*OLIG3-LOC100130476*	EA	LAN
2017	8	rs2955587	*FAM86B3P*	EA	LAN
2017	8	rs1966115	*PKIA-ZC2HC1A*	EA	LAN
2017	17	rs930297	*GRB2*	EA	LAN
2018	1	rs1780813	*SMYD3*	EUR	JUL
2018	5	rs55849330	*ST85IA4*	EUR	JUL
2018	7	rs150518861	*LAT2*	EUR	JUL
2018	17	rs114038709	*ARHGAP27*	EUR	JUL
2018	17	rs36023980	*GRB2*	EUR	JUL
2018	X	rs13440883	*GPR173*	EC	ZHG

KEY: EUR=European, EC=European and Chinese, EA=European American, BEN=Bentham et al [4], MOR=Morris et al [43], LAN=Langefeld et al [63], JUL=Julia et al [64], ZHG=Zhang et al [65].

1.5 Summary

This section introduced foundation information in genetics along with details of the autoimmune disease SLE. Also described was the application of previous statistical methods and the process that is used to produce a GWAS involving genetic data. Furthermore, the motivation was explained for this thesis which uses the Bentham et al data using advanced statistical methods which are explained in the next section.

2 Statistical Methods

In this section, the methods that are employed to analyse the data described in the previous chapter are explained, starting with basic information on frequentist statistics involving hypothesis testing, statistical significance and a correction method for false positives. The next section reviews the foundation in Bayesian statistics, presenting what the distributions are and how they are put together to form a posterior distribution using Bayes theorem. Exchangeability is described leading into hierarchical models that are the focus of the methods presented in this thesis. Regularization techniques are discussed with the various methods that have been applied to the data including optimization algorithms. The section is completed with an examination of the four methods that were used to analyse the data.

2.1 Introduction

A key issue in biostatistics is the vast amounts of variables that are present in genetic datasets. When dealing with the applications of genomics, proteomics and transcriptomics data involving hundreds of thousands (or even millions) of parameters, it would be desirable to reduce the number of random variables in a study. High dimensional data often have the number of features exceeding the number of observations with more predictors than data points (with large p and small n). Reduction in dimensionality is required ultimately reducing computational processing speeds. Reducing high dimensional data also helps interpretability, and relieves potential problems of overfitting. Another problem is the effects of linkage disequilibrium that produce multicollinearity amongst SNPs. Methods have been developed to account for all these problems including methods like variable selection and regularization.

2.2 Frequentist Statistics

2.2.1 Statistical Testing and Significance

Medical studies using the case-control method involve people who have the associated disease (cases) and a group that have no known link to the disease (the controls). They are run as hypothesis tests with the null hypothesis (H_0) representing no difference in allele frequency between the populations for each SNP while the alternative hypothesis (H_1) assumes that there is a difference between the populations. These tests are based on normal distributions and the usual level for significance testing is at the 5% level. In medical statistics rejecting the null hypothesis when true can have serious consequences. If one rejects the H_0 when it is true, this is classed as a Type I error with the probability of this represented by α. If one doesn't reject the H_0 when it is false, this is classed as a Type II error and the probability of this is represented by β. Optimally the test would have minimal values for α and β, where α is the SIZE of the test and $1 - \beta$ is the POWER of the test. When the null hypothesis of no difference is true (based on a 5% level) and is in the 95% confidence level

this is classed as non-significant. Also if the null hypothesis is not true and lies outside the confidence levels, this is classed as significant as shown in Table 5. Controlling these false positive associations is strongly desirable. Increasing the sample size also increases the power but at a cost to the size of the risk expected.

Table 5: Classification of Statistical Errors

	NON-SIGNIFICANT	SIGNIFICANT
TRUE H_0	CORRECT	Type I error
NOT-TRUE H_0	Type II error	CORRECT

Example:
A study of 1,000 SNPs at 5% significance level, resulting in 50 false positives
A study of 10,000 SNPs at 5% significance level, resulting in 500 false positives
A study of 100,000 SNPs at 5% significance level, resulting in 5000 false positives.
With hundreds of thousands of statistical tests per study, the need to correct for multiple comparisons is paramount to control false positive rates. The paper by Benjamini and Hochberg [66] entitled "Controlling the false discovery rate" which is a highly cited statistical paper, addressed the problem of multiple comparisons.
Generally, in genetics, when testing large amounts of data, studies now have 100,000 or more hypotheses, and it is clear to see that the potential resulting false positives are far too high. The resulting rejection area (say 5%) becomes sizeable and the chance of a rare event increases, thus a more stringent test is required. A potential way of solving this issue is to use the most widely used method, the Bonferroni Correction [67].
In the Bonferroni correction method, if there are m number of hypothesis tests with a rejection level of α, then we would only reject $\frac{\alpha}{m}$ tests. A standard GWAS threshold of $p = 5 \times 10^{-8}$ was used for association analysis in most GWA studies.

This method has been called into question from Perneger [68]. Perneger noted the inflation of type II error if the type I error decreases, debating that type I errors are no more false than type II. Also, Nakagawa [69] critiqued the correction method in a report noting a weakness in statistical power when rejecting an incorrect null hypothesis with a potential of publication bias to follow.

2.3 Bayesian Statistics

Bayesian statistics can be described as the technique of assigning probabilities given recorded data and updating previous views about unknown parameters to reach a suitable probability statement for complex problems. We treat the unknown parameters as random variables. The whole of Bayesian statistics is based around three probability distributions, the prior, the likelihood and the posterior.

The prior is a probability distribution that consists of prior beliefs about the true value of a parameter. These can be informative, where we have some kind of information that leads us to a certain distribution, weakly informative, where we have an approximate idea as to what the distribution could look like or uninformative, that relies on no information about the parameter. Informative and weakly informative priors allow some subjectiveness into the distribution chosen whereas uninformative priors lack any subjectivity. Whatever probability distribution is chosen for the prior, this is represented as $p(\theta)$.

The likelihood contains all the observed data that has been collected up to the current point for a fixed value y. This is represented as $p(y|\theta)$.

The posterior probability is found by using data that has been observed combined with subjective (or uninformative) prior beliefs of what the data could be. This produces an updated probability distribution that can be inferred from, to make an informed decision. It brings together an educated belief with known data to make a robust conclusion of unknown parameters. This is represented by $p(\theta|y)$. The posterior distribution is a combination of prior and likelihood resulting in the Equation 6.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int \{p(y|\theta)p(\theta)\} d\theta} \tag{6}$$

The posterior distribution is attained from Bayes' rule, the denominator is averaged over and returns a constant to produce Equation 7. With this new information our beliefs are changed, and we can infer from the posterior distribution for a more informed estimation.

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{7}$$

2.3.1 Exchangeability

For a sequence of random variables, $(y_1, y_2, ..., y_n)$, that are identically distributed with a symmetry for all permutations π of $1, ..., n$ and thus the product of Equation 7 does not depend on order, then in Equations 8 and 9 the π represents any permutation of the indices.

$$p(y_1, y_2, ..., y_n) = p(y_{\pi_{(1)}}, ..., y_{\pi_{(n)}}) \tag{8}$$

If Equation 8 is satisfied for any permutation π, it can be said that the uncertainty in the joint probability density is exchangeable. Based around the property of de Finetti's theorem of exchangeable sequences of random variables, exchangeability can be assumed in a joint density if there is a lack of knowledge about the random variables and they are conditionally independent [70]. Using the prior $p(\theta)$ and a sampling model $p(y_{\pi_{(i)}}|\theta)$

$$p(y_1, y_2, ..., y_n) = \int \Big\{ \prod_{i=1}^{n} p(y_{\pi_{(i)}}|\theta) \Big\} p(\theta) d\theta \tag{9}$$

We have independent samples from a population for a fixed but unknown value of a parameter ϕ.

If the observations are denoted as exchangeable, then a subset can be classed as a random sample of a model and thus a prior distribution on the parameters exists. This leads to an approach that requires Bayesian techniques that builds a framework with a hierarchical feature,

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi) \tag{10}$$

$$p(\theta) = \int \Big\{ \prod_{j=1}^{J} p(\theta_j|\phi) \Big\} p(\phi) d\phi \tag{11}$$

Due to ϕ being unknown, it can be averaged over the prior which integrates out the ϕ.

2.3.2 Bayesian Hierarchical Models

Hierarchical models are a natural way of taking into account relations between variables, by assuming a common distribution for a set of relevant parameters, thought to underlay the outcomes of interest The key advantage of the hierarchical approach is that it uses information across groups of observations to reduce our lower-level parameters' sensitivity to noise. A hierarchical model is a model in which lower levels are sorted under a hierarchy of successively higher-level units. It is often useful to think of the analysis of marketing data using one model for within-unit analysis, and another model for across-unit analysis. The within-unit model could be used to describe the behaviour of individual respondents over time, while the across-unit analysis could be used to describe the

diversity, or heterogeneity, of the units. The sub-models combine to form the hierarchical model, and Bayes theorem is used to integrate the pieces together and account for all the uncertainty that is present.

A hierarchical model is made from different sub-models, that model the relationships between variables and are related in some way ultimately producing a joint probability model. A hierarchical approach utilizes layered complex models that pool the information from separate groups that are not independent. This works well with haplotypes that have correlated SNPs and can be treated as dependent variables amongst each group.

Sophisticated models like Bayesian hierarchical models have seen steady growth in usage over recent years as computational power has increased. These models use two or more structured levels from multiple sources to combine to form the hierarchical model, using a tractable Bayesian prior with a well estimated likelihood to produce a posterior that can be easily sampled from. This technique helps with multiparameter problems and accounts for the heterogeneity of means across groups. The prior distribution parameters are known as hyperparameters.

Figure 4: A Bayesian Hierarchical model.

$$
\begin{array}{ccccc}
 & & \lambda_1, \lambda_2 & & \\
 & & \swarrow\downarrow\searrow & & \\
\theta_1 & & \theta_2 & & .....\theta_n \\
\downarrow & & \downarrow & & ..... \downarrow \\
Y_1 & & Y_2 & & ...Y_n
\end{array}
$$

Equation 12 shows a general Bayesian Hierarchical model with 2 levels.

$$P(\theta, \phi|Y) \propto P(Y|\theta)P(\theta|\phi)P(\phi) \tag{12}$$

The likelihood $P(Y|\theta)$ only depends on ϕ through θ. The distribution of the hyperprior is $P(\phi)$ and represents information about an unknown parameter. The joint posterior distribution $P(\theta, \phi|Y)$ is now a hierarchical model.

A 3 level hierarchical model contains a prior distribution, within group sampling variability and between group sampling variability. Equation 13 shows a general Bayesian Hierarchical model with 3 levels.

$$P(\theta, \phi, X|Y) \propto P(Y|\theta)P(\theta|\phi)P(\phi|X)P(X) \tag{13}$$

2.3.3 Bayes Factor

Bayes Factor (BF) is the measure of support of evidence to an association (in this study with SLE). Using multiple tests of different null hypotheses we can compare models resulting in Equation 14.

$$\text{BF} = \frac{\text{Likelihood of data given alternative hypothesis}}{\text{Likelihood of data given null hypothesis}} = \frac{P(y|H_1)}{P(y|H_0)} \tag{14}$$

The alternative hypothesis (H_1) is the probability of genotype configuration of association with SLE and the null hypothesis (H_0) being the probability of genotype configuration with genotype independence with SLE.

$$\text{BF} = \frac{\text{Posterior odds}}{\text{Prior odds}} \tag{15}$$

A BF of 1 shows no evidence to support the null hypothesis, meanwhile an increasing BF shows stronger support for the alternative hypothesis. A BF of less than one shows moderate eveidence to H_0 and a declining value to 0 produces stronger support to H_0.

2.3.4 Sampling from the Posterior Distribution

When inference is required from the posterior distribution, posterior sampling can be drawn by Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling [71] and the Metropolis Hastings algorithm [72]. Bayesian computation is also often sampled by means of numerical iterative algorithms. Techniques that evaluate gradients are popular, iterating until convergence to a maximum or a minimum with a specified stopping criterion. More recent work has seen the posterior sampled by numerical approximation methods including Expectation Propagation [73], Variational Bayes [74], and Expectation Maximization [75] techniques. The methods in this thesis will only use Expectation Maximization and numerical iterative algorithms.

2.4 Regularized Regression (Penalized Regression)

When using linear regression, we use ordinary least squares with the aim of minimizing the sum of squared errors. From the standard model for multiple linear regression, $Y = X\beta + \epsilon$, where Y is the response variable, the matrix X is made up of n observations x p predictors $(x_1.....x_p)$, β is a vector of regression coefficents where β_0 is the intercept and ϵ is an error term. When models are complex and have too many parameters to produce a reasonable model, a simplified model is required that retains most of the important information, using fewer parameters than a saturated model. This method of summarising parameters using sparsity can be adopted for large scale models as are needed for genetic datasets. With fewer variables in the model rather than a saturated one, the aim is to fit the best model for more accurate predictions producing a more efficient model. Penalized regression is required for interpretation, overfitting and underfitting problems and helps with reducing computational processing times.

For statistical models that have more predictors than observations using penalized regression methods can achieve superior fits resulting in an optimal model. A reduction in variance can be achieved but with a trade off with an increase in the bias. There have been many ways to attempt this including principal component analysis [76], backwards elimination, and forward selection [41] with many more subset selection algorithms that have been implemented. One needs to apply these methods when attempting to reduce dimensionality although correlated alleles make the process problematic. To address the problem of multicollinearity, variable selection and regularization can be employed to treat correlated variables with high "R scores" of linkage disequilibrium, by grouping all SNPs that are highly correlated with each other and producing just one predictor for the final model. Many early GWAS had not put in place this statistical advancement, that reduces the number of markers into a sparser more parsimonious model. The need to utilise variable selection to choose the relevant predictors that are in highly correlated haplotype blocks with linkage disequilibrium is required. To control overfitting which was previously performed by stepwise selection, penalized regression can manipulate the objective (or cost) function with the use of a penalty function.

2.4.1 Regularized Logistic Regression

In this GWAS, the aim is to find which predictors (SNPs) are important for associations with SLE. The response variable is binary, representing the presence of SLE or not, resulting in categorical data that is valued either as a 1 or a 0. The approach is to model the probability as a logistic model using the conditional mean of Y given x.

$$\pi(x) = E(Y|x) \tag{16}$$

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + x\beta \tag{17}$$

Solving for p , gives

$$P(y_i = 1) = \pi_i = \frac{e_i^x \beta}{1 + e_i^x \beta} \tag{18}$$

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \\ &= \sum_{i=1}^{n} y_i \log \left\{ \frac{\pi_i}{1 - \pi_i} \right\} + \log(1 - \pi_i) = \sum_{i=1}^{n} y_i x_i \beta - \log(1 + e^{x_i \beta}) \end{aligned} \tag{19}$$

The log likelihood function (Equation 19) is maximized for parameter estimation.

2.4.2 Regularization - Logisitc Ridge Regression

The earliest of the techniques to appear was ridge regression [77]. This method shrinks some of the coefficients more when the penalty function is larger, hence more of the coefficients are penalized. This method will keep all the coefficients in the model. The maximized log likelihood functions coefficients has a penalized parameter applied (barring the intercept) so Equation 19 now becomes logistic ridge regression as shown in Equation 20.

$$l(\beta) = \sum_{i=1}^{n} y_i x_i \beta - \log(1 + e^{x_i \beta}) - \lambda \sum_{j=1}^{p} \beta_j^2 \tag{20}$$

The ridge penalty function part is a squared value, and is constrained using a tuning parameter (λ) where $\lambda \geq 0$. This produces the l_2 penalties. The larger the value of λ the stronger the penalty is. The log likelihood function is then maximized and the coefficients will shrink towards zero (but not to exactly zero) and will remain in the model selection.

2.4.3 Regularization - Logistic Lasso Regression

From the origins of Robert Tibshirani's groundbreaking paper in 1996 [78], this method produced the choice to select a regressor and either keep it in the model or shrink its coefficient to zero, effectively eliminating its presence. This keeps the important covariates in the model that explain the data the best, creating a parsimonious model that results in a less complex modelling problem. This variable selection method was called least absolute shrinkage and selection operator, and the lasso was born. It has been noted in previous literature that the lasso does not handle correlated predictors well [79].

$$l(\beta) = \sum_{i=1}^{n} y_i x_i \beta - log(1 + e^{x_i \beta}) - \lambda \sum_{j=1}^{p} |\beta_j| \tag{21}$$

The penalty function part is an absolute value, and this produces the l_1 penalties as convex so numerical solutions can be found by coordinate descent techniques which are computationally efficient.

Many lasso methods have evolved since 1996 and have been adapted to produce many other variations: the group lasso in 2006 [80], the Bayesian lasso in 2008 [81], adaptive lasso in 2006 [82], and other techniques have arisen from mixing the lasso with another method: for example, the blending of the Spike and Slab Bayesian method with lasso in 2018 [83].

2.4.4 Regularization - The Elastic Net

In 2005, Zou and Hastie [79] devised an amalgamation of the ridge regression and the lasso method entitled "The Elastic Net". Zou and Hastie came up with the name as they thought "it is like a stretchable fishing net that retains all the big fish". Combining the l_1 and the l_2 penalties this method contains the ability to use variable selection while selecting groups of correlated variables. This grouping effect allows either the correlated group to be in the model or left out.

$$l(\beta) = \sum_{i=1}^{n} y_i x_i \beta - \log(1 + e^{x_i \beta}) + \alpha \sum_{j=1}^{p} \beta_j^2 + (1 - \alpha) \sum_{j=1}^{p} |\beta_j| \qquad (22)$$

The α is another tuning parameter. Once the predictors have been standardized and centred with variance of 1, the intercept (β_0) can be omitted. The left part is the mean squared error loss function, and the centre and right part is the regularization with penalty. When α is 1 the regularization becomes the lasso (l_1 norm). When α is 0 the regularization becomes ridge regression (l_2 norm). Solving the elastic net can be managed by gradient ascent and descent algorithms due to the convexity nature when $\alpha < 1$.

2.4.5 Optimization for regularization

Using regularization methods with certain parameters results in the need for convex optimization. Optimization aims to achieve high prediction accuracy by maximizing (for coordinate ascent) or minimizing (for coordinate descent). It is an iterative optimization algorithm to find the minimum of a function. Three of the statistical methods in this study use a form of pathwise coordinate optimization (EBEN [84] uses cyclical coordinate ascent, Lasso [85] and Spike [86] use cyclical coordinate descent). This technique searches for the complete set of solutions for the λ value (tuning parameter) then steadily increasing or decreasing the λ value until convergence. Using warm starts from the previous calculation this is a very efficient method. Most of the coefficients will be zero and the optimization exploits this sparsity.

In the case of logistic regression, a local search optimization algorithm is commonly used.

2.4.6 Expectation Maximization Algorithm

The expectation maximization algorithm (EM) is an iterative algorithm that is made up of two parts, the expectation step and the maximization step, starting with estimating the initial values of the parameters, then iteratively step by step updating the estimates until convergence. Originally an idea based on finding missing data problems, the EM is also used for maximum likelihood estimations, by means of numerical iterative computation, in frequentist statistics. In Bayesian statistics, many high dimensional problems become intractable when making approximations to the joint mode and they require the need for optimal approximations via the conditional posterior distribution. In 1977, Dempster, Laird and Rubin [75] published a paper showing that if the likelihood function is unimodal then the EM will guarantee to converge to that stationary point which is the Maximum Likelihood Estimator. Dempster, Laird and Rubin demonstrated that at no point after an EM iteration does the likelihood function decrease and thus showing monotonicity. In multimodal situations, it will converge to either a local maximum, a global maximum, or a saddle point (except exceptional cases) depending on what the first estimations were. In 1983 Wu [87], found an error in the proof of the EM and corrected it.

There are many numerical methods that can be implemented into problems to achieve a result. In this thesis there are optimization methods that entwine with others (e.g., Expectation Maximization Coordinate Descent Algorithm) are used.

2.4.7 Regularization - Variations

Park and Cassella (2008) [81] produced a "Bayesian Lasso", a lasso in Bayesian style with a hierarchical structure that treats the parameters as random variables. The Laplacian prior (also known as the double exponential prior) was the base for the conditional prior for β

$$\pi(\beta \backslash \sigma^2) = \sum_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma}^2} e^{-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}} \tag{23}$$

2.5 Frequentist Penalised Methods

The two frequentist methods differ by a lasso regularization technique. The paper by Kohannim et al [85] employed the lasso to produce a sparse but parsimonious fitted model that predicts disease association. The frequentist method by Lu et al [88] uses a standard statistical method for a GWAS and does not use variable selection. It is the only method in this study to leave all the predictors in the final model, whereas the other three methods use some type of regularization.

2.5.1 Frequentist approach with no regularization (Frequentist).

Based on the method in the paper by Lu et al [88], this standard method of multiple logistic regressions was processed by SNPTest software [12]. Quality control steps were conducted to improve the quality of the analysis, while missing information from individuals were removed from the GWAS [89]. These exclusions are SNPs that are missing from a large proportion of subjects. This is known as SNP-level missingness. After filtering, SNPs with a threshold rate of p <0.05 were kept in for analysis. Also, minor allele frequencies of less than 0.1 were discarded. Using an additive genetic model, the data was tested at each marker versus a model of no association. With no regularization this method kept all the predictors in and a vast amount of false positives was predicted.

2.5.2 Frequentist approach with Lasso regression (Lasso).

This analysis was based on the method in the paper by Kohannim et al [85]. In preparation for the method, the data was analysed by means of the software Plink [11]. The software was used to extract SNPs that had MAF greater than 0.1 and HWE p-values of less than 5.7 x 10^{-7}.
Using the penalized regression technique of lasso on the SNPs in a gene-centric manner, this method allows for correlated variables in the gene and produces sparse groups amongst the SNPs.

The standard lasso regression model is

$$\beta^* = argmin||y - X\beta||^2 + \lambda||\beta||_1 \tag{24}$$

where X is the matrix from the additive model of single SNPs, β^* is the vector of fitted coefficients, and λ is the tuning parameter of the penalty function. This was processed using the R package GLMnet [90].

It was required to find the optimal penalty parameter for the lasso regularization, and this was achieved through leave one out cross validation. In general, cross validation divides the sample data into two subsets, one of the subsets being the training data and the rest of the data being the validation set [91]. Successive rounds are repeatedly performed using different subsets of the training and validation data. The evaluated value from each round is combined to produce an estimate of the parameter that does not overfit or underfit the model. This prediction using computational means measures the combined fitness of the model. Meanwhile, leave one out Cross validation (When K=N) uses all but one of the data as training sets while leaving just a single datum as a test set. This can result in a drawback of low bias with large variance due to all training sets being near identical. This method is computationally expensive so Kohannim et al added cyclical coordinate descent in a path-wise fashion to the algorithm to increase speed of the process, producing greater confidence in the results, as described in the original paper.

Due to the function convexity of β we can utilize cyclical coordinate descent. The objective function is optimized while searching for the best direction keeping the other parameters at a fixed amount for each search, repeating cycle after

cycle searching for the best update at each iteration until this function converges to a global minimum after following a path that utilizes warm starts from each previous calculation. The full solution is calculated over the entire range of values for λ concluding in a fast efficient algorithm.

2.6 Bayesian Methods Applied

The two methods used here are Bayesian Hierarchical models, each with a different type of regularization. The first method is "Spike and Slab Lasso GLM" [86] from 2017. This advanced method contains multiple statistical techniques to achieve sparse regularization. The second method "The Empirical Bayesian Elastic Net" (EBEN) from 2015, incorporates the elastic net into a Bayesian environment which manipulates groups of strongly correlated variables into a single important non-zero coefficient [84]. They both feature coordinate optimization methods and a variant of cross validation.

2.6.1 Spike and Slab Lasso GLM (Spike).

This method was based on the paper by Tang et al [86] and is a blend of the lasso [78], Generalised linear models [92], Bayesian hierarchical models and the expectation maximization algorithm for numerical optimization [75]. It has a foundation in generalized linear models (GLM) with logistic regression. A GLM is an extension of a linear model using three components: a link function, a linear predictor, and a response distribution.

$$\eta_i = \beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j = X_i\beta \tag{25}$$

Here β is a vector of the intercept and all the coefficients, X_i is all variables. The relationship involves a link function

$$E(y_i|X_i) = h^{-1}(X_i\beta) \tag{26}$$

giving the response of a binomial distribution

$$p(y|X\beta,\phi) = \sum_{i=1}^{n} p(y_i|X_i\beta,\phi) \tag{27}$$

Tang et al [86] noted that "GLM's cannot jointly analyse multiple correlated predictors due to unidentifiability and overfitting". The GLM will need modifying with a form of regularisation. The lasso penalty penalizes the maximum log likelihood resulting in convexity that can be optimized through coordinate descent.

The roots of the spike and slab materialised from Lempers [93] and was carried forward by Mitchell and Beauchamp [94] when they first quoted the term "Spike and Slab". In 1993, George and McCulloch [95] came up with the term

Stochastic Search Variable Selection (SSVS) which is built through a hierarchical Bayesian framework. It is designed to detect the best subsets of variables from two normal mixture models. It produces strong shrinkage on irrelevant coefficients resulting in a mass around 0 (the spike) and a diffuse distribution to weakly shrunk coefficients (the slab). This is mixed with the generalized linear model that contains a large number of correlated predictors. Starting with a Bayesian Hierarchical modelling framework, the Laplace distribution (also known as the double exponential distribution (DE)) serves as a suitable model for the prior (Equation 28) which places mass at 0 (spike) or on large values and thus in the tails (slab) [78], where the s controls the amount of shrinkage.

$$\beta_j|s \sim DE(\beta_j|0,s) = \frac{1}{2s}\exp\left(-\frac{|\beta_j|}{s}\right) \tag{28}$$

The spike and slab mixture double exponential prior is shown in Equation 29.

$$\beta_j|\gamma_j, s_0, s_1 \sim (1-\gamma_j)DE(\beta_j|0,s_0) + \gamma_j DE(\beta_j|0,s_1) \tag{29}$$

The indicator Variables (γ_j), are γ_j =1 or 0, and the scale (S_j) is $(1-\gamma_j)s_0$ + $\gamma_j s_1$. The scale s_0 is chosen to be small and acts as the spike, hence strong shrinkage. The scale s_1 is chosen to be large and acts as the slab, hence weak or no shrinkage. Equation 30 shows the link that the scale parameters have with the coefficients where θ is the probability parameter and assume that $\theta \sim U(0,1)$.

$$\gamma_j|\theta \sim Bin(\gamma_j|1,\theta) = \theta^{\gamma_j}(1-\theta)^{1-\gamma_j} \tag{30}$$

The EM algorithm (described above) finds the marginal posterior modes in a Bayesian context. In the paper by Tang et al [86], the EM algorithm is mixed with a coordinate decent algorithm treating the indicator variables γ_j as missing values. The development of a fast efficient algorithm was required to fit the spike and slab lasso GLM important predictors.

EM Coordinate Descent Algorithm

The EM coordinate descent algorithm is based on the log joint posterior density of the parameters $(\beta, \phi, \gamma, \theta)$

$$\log p(\beta,\phi,\gamma,\theta|y) = \log p(y|\beta,\phi) + \sum_{j=1}^{J}\log p(\beta_j|S_j) + \sum_{j=1}^{J}\log p(\gamma_j|\theta) + \log p(\theta) \tag{31}$$

$$\propto l(\beta,\phi) - \sum_{j=1}^{J}\frac{1}{S_j}|\beta_j| + \sum_{j=1}^{J}[(\gamma_j\log\theta + (1-\gamma_j)\log(1-\theta)] \tag{32}$$

The maximizing step is obtained from the log joint posterior (β, ϕ).

$$l(\beta,\phi) - \sum_{j=1}^{J}(S_j^{-1}|\beta_j|) \tag{33}$$

They are updated by maximizing $Q_1(\beta, \phi)$ with the second part serving as the L_1 lasso penalty.
From the log joint posterior θ is updated by maximizing the expressions Q_1 and Q_2.

$$Q_1(\beta, \phi) = l(\beta, \phi) - \sum_{j=1}^{J} \frac{1}{s_j} |\beta_j| \tag{34}$$

$$Q_2(\theta) = \sum_{j=1}^{J} [\gamma_j \log \theta + (1 - \gamma_j) \log(1 - \theta)] \tag{35}$$

The expectation step involves updating γ_j and S_j^{-1} by their conditional posterior expectations.

2.6.2 Empirical Bayesian Elastic Net Method (EBEN).

This method is based on the paper by Huang, Xu and Cai [84]. Huang, Xu and Cai had presented their earlier studies on their EBlasso algorithms [96, 97] that laid the foundations for their 2015 paper. Their work was based on Bayesian hierarchical models with Normal, Gamma and Exponential priors. The Empirical Bayesian Elastic Net method (EBEN) incorporates the elastic net into a Bayesian environment. Huang, Xu and Cai developed the EBEN that manipulates groups of strongly correlated variables into a single important non-zero coefficient. This method applies strict variable selection. The EBlasso has an exponential prior distribution for the variance components but could not reveal the highly correlated predictors. The method uses a non-informative uniform prior with unknown parameters μ, $p(\mu) \propto 1$ and σ_0^2, $p(\sigma_0^2) \propto 1$ using a two level hierarchical model for β to help with correlated variables. The first level follows a normal distribution $\beta_j \sim N(0, \sigma_j^2)$, with

$$\alpha_j = \frac{1}{\sigma_j^2} \tag{36}$$

The second level is a Gamma distribution (Equation 38) where λ_1 and λ_2 are the hyperparameters, and α_j is decomposed as $\lambda_1 \geq 0$ and $\tilde{\alpha_j} > 0$ in Equation 37.

$$\alpha_j = \lambda_1 + (\tilde{\alpha_j}) \tag{37}$$

$$f(\sigma_j^2) = c(\lambda_1 \sigma_j^2 + 1)^{-(\frac{1}{2})} \exp(-\lambda_2 \sigma_j^2) \tag{38}$$

The prior distribution of β when the value of $\lambda_1 = 0$ produces the exponential distribution $p(\beta_j)$ (Equation 39) which is the Laplace distribution using the lasso penalty.

$$p(\beta_j) \propto \exp(-\sqrt{2\lambda_2}|\beta_j|) \tag{39}$$

$$f(\sigma_j^2|a,b,\gamma) = \frac{b^a}{\Gamma(a)}(\sigma_j^2 - \gamma)^{a-1}\exp(-b(\sigma_j^2 - \gamma)) \tag{40}$$

The prior distribution of β when $\lambda_1 > 0$, $\lambda_2 \geq 0$, $a = \frac{1}{2}, b = \lambda_2$ and $\gamma = -\frac{1}{\lambda_1}$and c=$\sqrt{\lambda_1, \lambda_2/\pi}exp(-\lambda_2/\lambda_1)$ becomes a shifted Gamma distributiom (Equation 40) which uses the elastic net penalty, as shown in Equation 41.

$$p(\beta_j) \propto \exp\left(-\frac{\lambda_1}{2}\beta_j^2 - \sqrt{2\lambda_2}|\beta_j|\right) \tag{41}$$

To control the degree of shrinkage, two hyperparameters (λ_1 and λ_2) were derived by means of 5-fold cross validation which obtained the optimal parameters for input into the EBEN algorithm. When we require a measurement of estimation for prediction error, a common method is cross validation. To combat overfitting in a model, cross validation can optimize the prediction of new data using K-fold cross validation.

The data is divided into subsets of K, usually 5 or 10, with $K-1$ being the training sets and one subset being the test set. There are no overlapping sets, and all the data is used up in either type of set. The validation results are then averaged and the prediction error is calculated for the $K-1$ sets. Table 6 shows an example of 5-Fold cross validation. Each row is an iteration of the full data set

Table 6: 5-Fold cross validation.

An example of 5-Fold cross validation				
Test	Tr	Tr	Tr	Tr
Tr	Test	Tr	Tr	Tr
Tr	Tr	Test	Tr	Tr
Tr	Tr	Tr	Test	Tr
Tr	Tr	Tr	Tr	Test

Cross Validation used in EBEN, where $\nu \in [0,1]$:

$$\lambda_1 = (1-\nu)\lambda$$

and

$$\lambda_2 = \nu\lambda$$

The smallest pair of (ν, λ) from the prediction error calculations are taken as the

optimal parameters, with prediction error $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$. The posterior distribution of β is inferred after the unknown parameters are estimated.

The posterior distribution of parameters θ, where $\theta = (\mu, \sigma_0^2, \beta, \tilde{\sigma}^2)$ is given by:

$$p(\theta|y) \propto p(y|\mu, \beta, \sigma_0^2)p(\mu)p(\sigma_0^2)p(\beta|\sigma^2)p(\sigma^2|\lambda_1\lambda_2) \tag{42}$$

The marginal posterior distribution has the β integrated out

$$p(\mu, \tilde{\sigma}^2, \sigma_0^2|y) = \int p(\theta|y)d\beta \tag{43}$$

The likelihood $P(y|\mu, \beta, \sigma_0^2)$ only depends on λ through σ^2.
The coordinate ascent method is used to estimate the parameters $\alpha_1, ..., \alpha_k, \mu$ and σ_0^2. The log marginal posterior distribution has a global maximum. This is found through the EBEN algorithm involving coordinate ascent numerical method (iterations until convergence) and concludes when it achieves a specified value. The optimal value for α_j^* maximizes the log marginal posterior distribution solved with the following equations:

$$\alpha_j^* = \begin{cases} r, \text{if } q_j^2 - s_j > \lambda_1 + 2\lambda_2, \infty, \text{otherwise} \end{cases} \tag{44}$$

where

$$s_j = x_j^T C_{-j}^{-1} x_j, q_j = x_j^T C_{-j}^{-1}(y - \mu) \tag{45}$$

$$r = -\frac{s_j + \lambda_1 + 4\lambda_2) - \sqrt{\Delta}}{2(s_j - q_j^2 + \lambda_1 + 2\lambda_2)}.(s_j + \lambda_1) \tag{46}$$

with also

$$\Delta = (s_j + \lambda_1)^2 + 8\lambda_2 q_j^2 \tag{47}$$

$$C = \sigma_0^2 I + \Sigma_{j=1}^{k}(\lambda_1 + \tilde{\alpha_j})^{-1} x_j x_j^T \tag{48}$$

$$C_{-j} = C - (\lambda_1 + \tilde{\alpha_j})^{-1} x_j x_j^T \tag{49}$$

2.7 Applied Methods Summary

The previous section has explained the theory of all four methods applied to Bentham et al data [4]. These are two types of frequentist methods and two Bayesian techniques. The frequentist method with no variable selection, will keep all variables in the model, in contrast to the more sophisticated method, the lasso, which will reduce the predictors. The key research questions are whether these methods will find SNPs known to be associated with lupus, whether the methods will produce large numbers of false positive results, and ultimately whether these complex variable selection methods can be used as an alternative to the current simplistic statistical models used in GWA studies.

3 Results

The results section is divided into seven parts. The six chromosomes with the most well established associations appear after an initial review of the whole genome. For each chromosome there is a short introduction reviewing the reported pre-GWAS associations with SLE and a timeline including a review of post GWAS findings. Every entry featured, has been associated with SLE pertaining to a partial or full dataset of European ancestry so a comparison can be made with this study's European based data set. Each subsection presents the results of each chromosome by means of comparison tables, Manhattan plots and a linkage disequilibrium plot that involves a relevant block of SNPs. For each method, the SNPs are ranked, analysed and comparisons are made with previous known associations. The remaining results that have been produced arise in the appendix, for each of the other autosomal chromosomes

3.1 Whole Genome

3.1.1 Introduction

The post-cleaned data from the Bentham et al study [4] was analysed using the four different statistical methods described previously, to produce the results that will appear presented in tables, and graphically in Manhattan plots. There will be two types of Manhattan plot that will appear in this thesis depending on which methods data it is supporting. For the frequentist and EBEN methods the plots will have a y-axis of negative log values representing p-values, meanwhile the lasso and spike methods Manhattan plots will show negative log beta coefficients on the y-axis. All the Manhattan plots present the results in a graphic form with the chromosomes 1 - 22 on the x-axis. Each data point represents a SNP that has been analysed using one of the four methods and any data points (representing p-values) that feature higher than the red genome wide significance value of $-\log_{10}(5.00E-08)$ are classed as significant. The blue line is measured at $-\log_{10}(1.00E-05)$ as the commonly used genome wide suggestive association value.

Also featured graphically are linkage disequilibrium (LD) plots. These plots will concentrate on a small area of SNPs that are close together, surrounding a previously associated SNP or a locus of interest. The LD plots demonstrate the correlation between neighbouring SNPs and the possibility of making associations with disease and non-causal marker.

Three methods processed the large data set approximately at the same speed but the computation time of the EBEN method was considerably longer than the other three methods.

3.1.2 Associations with SLE

The published associated SNPs that arose from the results of the original study by Bentham et al [4], are compared with the Spike method and the Lasso method in Table 7, and are also compared with the Frequentist method and the EBEN

method in Table 8. The SNPs rs9652601, rs34572943 and rs2286672 were imputed in the original study and so do not appear in this analysis. If a zero appears in the coefficient columns, this equates to a variable selection method not selecting this SNP to be part of the model, and thus no ranking is made. From Table 7 the spike method produced 14 zero beta-coefficients with just one associated SNP in the top 25 rankings and six in the top 100. The lasso method calculated 10 zero beta-coefficients and also had only one SNP in the top 25 and eight in the top 100.

Table 8 shows the EBEN method recorded 11 zero beta coefficients with 10 top 25 ranking SNPs of associated risk alleles and fifteen in the top 100. The EBEN has chosen more of the previously associated SNPs than the other two variable selection methods in this area, although only one associated SNP has a genetically significant p-value by the EBEN method and that is rs7444 (2.52E-08). The frequentist method found 26 SNPs in the top 100 rankings, with 22 in the top 25. 25 of the associated SNPs were recorded as being statistically significant by the frequentist method. From the tables of top ten ranked SNPs by method that are highlighted throughout the results section by chromosome (see below), 32 of the 60 SNPs were shared between the spike and lasso method. The most extreme value from Bentham et al study was in chromosome 6. SNP rs1270942 recorded a p-value of 1.70E-101. In this study, the frequentist method also recorded an extreme value of 6.31E-81 for this SNP, although the variable selection methods all recorded zero beta coefficients. This marker has negligible linkage disequilibrium with close by SNPs although it is inside the major histocompatibility complex.

Table 7: Associated SNPs from Bentham et al

Associated SNPs						
Chr	SNP	Bentham p-value	SPIKE Coefficient	RANK	LASSO Coefficient	RANK
1	rs2476601	8.34E-13	3.31E-01	160	2.17E-01	48
1	rs1801274	6.05E-11	1.13E-01	168	9.01E-02	172
1	rs704840	1.65E-13	7.43E-02	189	7.49E-02	206
1	rs17849501	1.63E-59	-9.25E-01	17	-6.54E-01	2
1	rs3024505	2.55E-03	0	0	0	0
1	rs9782955	5.58E-04	0	0	3.39E-02	469
2	rs6740462	2.31E-08	9.90E-02	151	8.85E-02	158
2	rs2111485	3.44E-06	-3.79E-02	388	-3.16E-02	44
2	rs11889341	1.17E-65	-5.54E-01	71	-3.46E-01	20
2	rs3768792	2.35E-08	0	0	5.66E-02	257
3	rs9311676	5.37E-06	0	0	0	0
3	rs564799	1.15E-06	0	0	0	0
4	rs10028805	4.50E-10	0	0	0	0
5	rs7726414	9.17E-10	-4.52E-01	82	-2.63E-01	46
5	rs10036748	2.83E-18	-2.60E-01	139	-1.22E-01	167
5	rs2431697	3.23E-14	-1.14E-01	145	-1.30E-01	148
6	rs1270942	1.70E-101	0	0	0	0
6	rs9462027	1.80E-05	0	0	0	0
6	rs6568431	4.33E-12	2.27E-01	169	1.28E-01	145
6	rs6932056	1.23E-16	0	0	3.13E-02	793
7	rs849142	3.49E-05	5.27E-02	190	4.03E-02	519
7	rs4917014	4.10E-05	-5.46E-02	179	-5.78E-02	311
7	rs10488631	2.66E-44	2.97E-01	122	3.27E-01	28
8	rs2736340	2.14E-16	0	0	0	0
10	rs2663052	1.59E-08	1.03E-01	151	9.27E-02	256
10	rs4948496	1.17E-06	0	0	0	0
11	rs12802200	8.43E-09	-2.65E-01	124	-1.11E-01	100
11	rs2732549	1.31E-10	0	0	0	0
11	rs3794060	1.13E-04	5.21E-02	198	4.77E-02	302
11	rs7941765	9.82E-07	9.02E-02	127	5.67E-02	241
12	rs10774625	9.47E-08	1.57E-02	654	1.88E-02	893
12	rs1059312	3.20E-06	0	0	2.52E-02	730
14	rs4902562	4.85E-05	5.09E-02	204	4.57E-02	303
15	rs2289583	9.35E-09	1.64E-01	93	1.38E-01	66
16	rs11644034	1.25E-15	-1.99E-02	567	-4.08E-02	461
17	rs2941509	4.32E-06	0	0	0	0
19	rs2304256	2.34E-12	-1.43E-01	65	-6.52E-02	942
22	rs7444	1.30E-13	1.03E-01	76	1.45E-01	485

Table 8: Associated SNPs from Bentham et al

Associated SNPs						
Chr	SNP	Bentham p-value	FREQ p-value	RANK	EBEN p-value	RANK
1	rs2476601	8.34E-13	3.20E-10	25	3.08E-04	4
1	rs1801274	6.05E-11	1.51E-09	28	1.82E-03	19
1	rs704840	1.65E-13	5.69E-11	18	1.03E-01	1910
1	rs17849501	1.63E-59	1.50E-72	1	0	0
1	rs3024505	2.55E-03	1.58E-02	4051	0	0
1	rs9782955	5.58E-04	6.03E-05	262	3.07E-01	1227
2	rs6740462	2.31E-08	9.26E-09	32	2.33E-02	364
2	rs2111485	3.44E-06	6.46E-06	155	5.24E-02	601
2	rs11889341	1.17E-65	1.37E-70	2	1.35E-04	40
2	rs3768792	2.35E-08	4.96E-08	46	1.24E-01	1100
3	rs9311676	5.37E-06	2.35E-03	1159	0	0
3	rs564799	1.15E-06	4.28E-11	9	2.23E-04	994
4	rs10028805	4.50E-10	1.18E-08	12	0	0
5	rs7726414	9.17E-10	4.66E-17	5	3.70E-01	306
5	rs10036748	2.83E-18	3.30E-22	2	7.69E-03	5
5	rs2431697	3.23E-14	3.82E-14	8	3.83E-03	1
6	rs1270942	1.70E-101	6.31E-81	4	0	0
6	rs9462027	1.80E-05	1.36E-07	1658	0	0
6	rs6568431	4.33E-12	2.06E-11	1112	2.27E-05	4
6	rs6932056	1.23E-16	5.12E-16	704	0	0
7	rs849142	3.49E-05	4.28E-04	277	1.09E-01	64
7	rs4917014	4.10E-05	1.88E-04	181	8.13E-02	33
7	rs10488631	2.66E-44	8.86E-43	1	5.85E-03	6
8	rs2736340	2.14E-16	2.96E-15	4	0	0
10	rs2663052	1.59E-08	8.72E-05	187	4.43E-04	45
10	rs4948496	1.17E-06	1.47E-09	6	0	0
11	rs12802200	8.43E-09	3.05E-11	6	5.11E-02	314
11	rs2732549	1.31E-10	9.21E-13	3	0	0
11	rs3794060	1.13E-04	9.76E-10	18	4.32E-03	60
11	rs7941765	9.82E-07	6.57E-05	205	5.71E-04	17
12	rs10774625	9.47E-08	3.10E-08	14	1.10E-01	156
12	rs1059312	3.20E-06	1.93E-07	27	1.76E-01	111
14	rs4902562	4.85E-05	2.55E-02	2346	2.32E-02	266
15	rs2289583	9.35E-09	2.36E-09	9	9.18E-04	1
16	rs11644034	1.25E-15	1.17E-21	9	2.45E-01	778
17	rs2941509	4.32E-06	1.21E-05	36	0	0
19	rs2304256	2.34E-12	2.85E-14	1	9.65E-05	1
22	rs7444	1.30E-13	3.84E-13	2	2.52E-08	6

Figure 5 is drawn from the frequentist methods data, highlighting all SNPs from the whole genome (minus the sex chromosomes) that have been processed. Note the tall column of SNP values on chromosome 6 that is topped by the SNP rs2854275. This is the dense complex region of SNPs called the MHC.

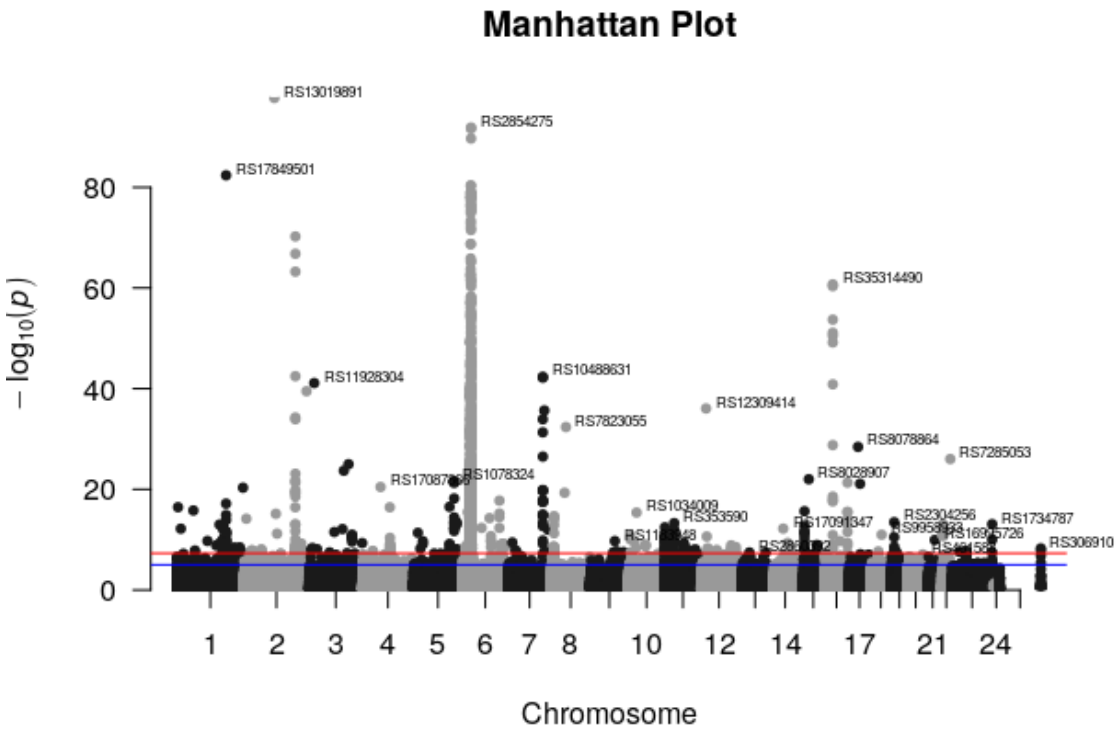

Figure 5: A Manhattan plot with a selection of the lowest p-values annotated

3.1.3 Non-Zero coefficient SNPs across all 4 methods

Across the genome, in general, it is noticeable from Table 9 and Figure 6 (below) that the number of non-zero coefficients chosen by each variable selection method per chromosome, increases proportionally, as the chromosomes get smaller in physical size. There is no obvious genetic reason for this and this could be an area for future work as this may be artefactual. There are a few cases where there is an extreme difference in numbers of non-zero coefficients chosen. For instance, the spike and the lasso method selected far fewer variables in chromosome 3 than the EBEN method, by a considerable amount. The EBEN produced 18.47% of SNPs with non-zero coefficients, meanwhile the spike and lasso recorded 8.98% and 7.52% respectively. This also occurred in chromosome 19 and 22 where this time it was the lasso method that chose far higher numbers of SNPs. The Lasso produced 25.42% for chromosome 19 , meanwhile the spike method produced 8.98% and the EBEN method 7.52%. For chromosome 22 the Lasso produced 33.9% non-zero coefficents, while the Spike method made 12.23% and the EBEN 11.62%.

Table 9: The amount and the percentage of SNPs with non-zero coefficient chosen by the spike, lasso and the EBEN methods. Figures in bold are extreme comparison percentages compared to the other methods per chromosome.

Percentage of SNPs with non-zero coefficient chosen by each method							
Chr	SNPs	SPIKE	%	LASSO	%	EBEN	%
1	52418	1801	3.44%	1701	3.25%	2304	4.40%
2	51086	1752	3.43%	1556	3.05%	3793	7.42%
3	42483	1443	3.40%	1645	3.87%	7848	**18.47%**
4	36441	1222	3.35%	1716	4.71%	970	2.66%
5	37924	1368	3.61%	2605	6.87%	850	2.24%
6	42993	1255	2.92%	2309	5.37%	488	1.14%
7	33932	1366	4.03%	2246	6.62%	1154	3.40%
8	33319	1328	3.99%	1511	4.53%	1820	5.46%
9	29764	1349	4.53%	2390	8.03%	1549	5.20%
10	34698	1469	4.23%	2413	6.95%	1551	4.47%
11	32336	1319	4.08%	1648	5.10%	788	2.44%
12	31785	1280	4.03%	1879	5.91%	1504	4.73%
13	24265	1175	4.84%	1470	6.06%	1334	5.50%
14	20959	1438	6.86%	1398	6.67%	1038	4.95%
15	19521	1143	5.86%	1196	6.13%	747	3.83%
16	20388	1272	6.24%	1772	8.69%	1303	6.39%
17	18084	1438	7.95%	1307	7.23%	717	3.96%
18	19037	1422	7.47%	2386	12.53%	1037	5.45%
19	13157	1181	8.98%	3345	**25.42%**	990	7.52%
20	16936	1171	6.91%	1101	6.50%	568	3.35%
21	9324	1124	12.05%	1556	16.69%	725	7.78%
22	9563	1170	12.23%	3242	**33.90%**	1111	11.62%

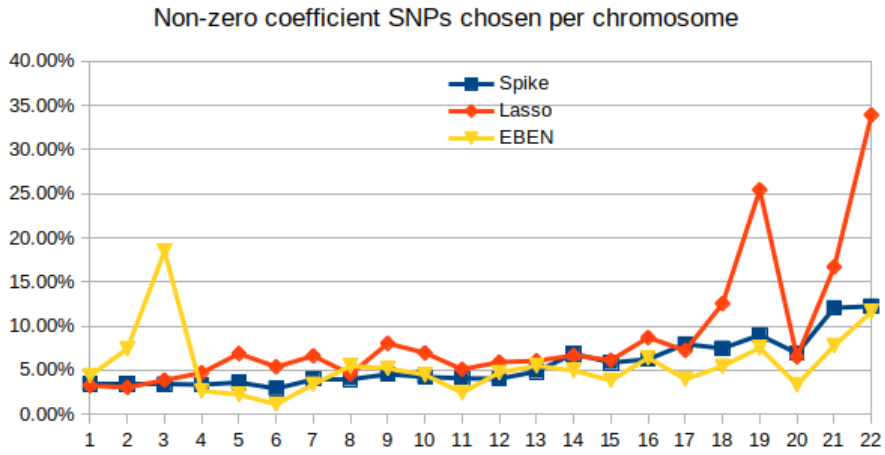

Figure 6: As the chromosomes decrease in physical size (from chromosome 1 to 22) the percentage of non-zero coefficient SNPs chosen increased.

3.1.4 SNPs that agree across all 4 methods

In Chromosome 2, SNP rs2573219 appears in the top 14 associations for all 4 methods. In chromosome 4, SNP rs17087866 appears third in the spike and slab method and had the strongest association for the other methods. This SNP has zero LD with nine SNPs either side. In chromosome 5, SNP rs1078324 was top in two methods and in the top 25 for the lasso and elastic net. In chromosome 7, SNP rs10264693 appears in the top 3 of all 4 methods. In chromosome 10, SNPs rs12775513 and rs10826385 were in the top 28 of all 4 methods. Both SNPs has zero LD with closeby SNPs. In chromosome 12, SNP rs1564363 featured in the top 15 of all 4 methods. In chromosome 14, SNP rs6575958 featured in the top 13 of all methods and had low LD with neighbouring SNPs. In chromosome 20, SNP rs6023308 was ranked first in the lasso and the elastic net and was also in the top 8 for the other 2 methods. In chromosome 22, SNP rs7285053 was ranked first by all 4 methods and was the only marker for which this was the case. A common theme across the genome is SNPs that are highly ranked (minimum 28th) across all four methods, demonstrate a lack of linkage disequilibrium with their neighbouring SNPs. All the SNPs that are included in this set also have had no previous reported associations to any disease, which suggests that they are false positives.

3.2 Chromosome 1

3.2.1 Introduction

Chromosome 1 is the largest chromosome in a human. Previously, there have been many studies that have produced associated SNPs with lupus on chromosome 1 resulting in this chromosome being of particular interest. The gene protein tyrosine phosphatase nonreceptor type 22 (*PTPN22*) [98] and the receptor type II (FcyRII) gene *FCGR2A* and FcyRIII gene *FCGR3A* [99] were found to have significant associations to lupus even before the GWAS era.

3.2.2 Previous GWAS associations in Chromosome 1

The first GWAS association with lupus was made in the gene *TNFSF4* by Harley et al [41]. In 2009, SNP rs3024505, that lies downstream of the *IL10* gene, which had previously been associated with multiple diseases such as inflammatory bowel disease, type I diabetes, ulcerative colitis, and Crohn's disease was also linked with lupus in a study of European and Asian populations by Gateva et al [60]. In the same study, Gateva et al found an associated SNP rs9782955 in the *LYST* gene and this was replicated by Bentham et al [4]. The SNP rs525410 (in or near *LAMC2*) was found to be associated with lupus by Chung et al in 2011 [61]. This study was based on the data from the Hom et al study although no association was found in the replication study based on the Harley et al data [7]. The Bentham et al study found SNP rs704840 to be associated with lupus after Martin et al [100] had reported this risk locus 2 years previously, with both studies based on European populations. SNP rs10911628 was associated with lupus after a study by Armstrong et al in a European population [62]. In 2016, Morris et al [43], produced a study combining European population (4,036 cases and 6,959 controls) with Chinese population (meta-analysis of two Chinese GWASs comprising 1,659 cases and 3,398 controls). Two associations were found at SNPs rs34889541 in or near gene *PTPRC* and rs2297550 in or near gene *IKBKE*. SNP rs6662618 in the gene region *GFI1-EVI5* has been associated with SLE in 2017 by Langefeld et al [63] and multiple sclerosis by Wang et al [101] in 2016. The SNP rs1780813 that is in the intron of the gene *SMYD3*, had been associated with lupus in the 2018 study by Julia et al [64]. Table 10 shows the associations with lupus from GWA studies that utilize partial or whole European populations over the period from 2008-2018.

Table 10: Timeline of associated SNPs with lupus in Chromosome 1

GWAS 2008-2018					
Year	Chr	SNP	Likely causal gene	Population	Author
2008	1	rs10798269	*TNFSF4*	EUR	HAR
2009	1	rs3024505	*IL10*	EA	GAT
2009	1	rs9782955	*LYST*	EA	GAT
2011	1	rs525410	*LAMC2*	EUR	CHU
2013	1	rs704840	*TNFSF4*	EUR	MAR
2014	1	rs10911628	*EDEM3*	EUR	ARM
2016	1	rs34889541	*PTPRC(CD45)*	EC	MOR
2016	1	rs2297550	*IKBKE*	EC	MOR
2017	1	rs6662618	*GFI1-EVI5*	EAH	LAN
2018	1	rs1780813	*SMYD3*	EUR	JUL

KEY: EUR=European, EC=European and Chinese, EA=European American, EAH=European, African and Hispanic Amerindian, ARM=Armstrong et al [62], CHU=Chung et al [61], GAT=Gateva et al [60], HAR=Harley et al [41], JUL=Julia et al [64], LAN=Langefeld et al [63], MAR=Martin et al [100], MOR=Morris et al [43].

3.2.3 Results - Spike and Slab Method

Figure 7 is from the data of the spike method and presents a Manhattan plot of SNP beta coefficients from chromosome 1. It highlights the SNP with the most significant coefficient on chromosome 1 from the spike method (rs1780813). This marker was ranked first by the lasso and the spike method with the frequentist method ranking it the second most significant.

Table 11: Top ten SNPs ranked for Chromosome 1 and accompanied with their coefficient

Spike and Slab			
SNP	Position	RANK	COEFFICIENT
RS1780813	246444082	1	-2.16E+00
RS6681218	47959645	2	1.38E+00
RS11808094	47989572	3	-1.35E+00
RS12061058	185891053	4	1.33E+00
RS17117228	98271048	5	1.29E+00
RS904295	100963041	6	1.23E+00
RS17020562	213542706	7	1.17E+00
RS6680879	100965683	8	1.12E+00
RS11204930	152151950	9	1.05E+00
RS10911789	185940266	10	-1.03E+00

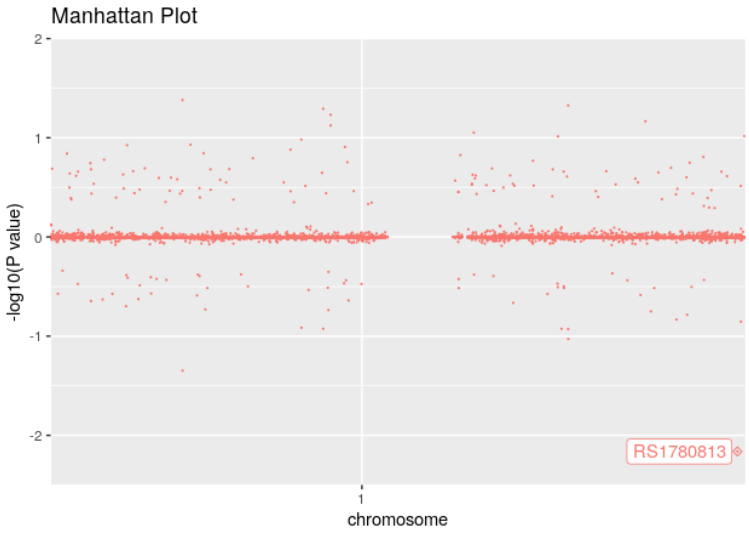

Figure 7: A Manhattan plot with the top SNP highlighted for the spike method

3.2.4 Results - Lasso Method

Table 12 presents the top ten ranked SNPs for chromosome 1 along with beta coefficients. Figure 8 shows a Manhattan plot with each beta coefficients as a point. This has been produced from the lasso data and is annotated with the top 5 extreme valued SNPs.

Table 12: Top ten SNPs ranked for Chromosome 1 and accompanied with their coefficient. The associated SNP is highlighted in bold.

Lasso			
SNP	POSITION	RANK	COEFFICIENT
RS1780813	246444082	1	-1.35E+00
RS17849501	**183542323**	**2**	**-0.65E+00**
RS35358165	248900100	3	0.60E+00
RS6681218	47959645	4	0.58E+00
RS17020562	213542706	5	0.57E+00
RS11204930	152151950	6	0.53E+00
RS17117228	98271048	7	0.48E+00
RS11811658	6641758	8	0.48E+00
RS2256917	182249873	9	0.46E+00
RS17117203	98265210	10	-0.44E+00

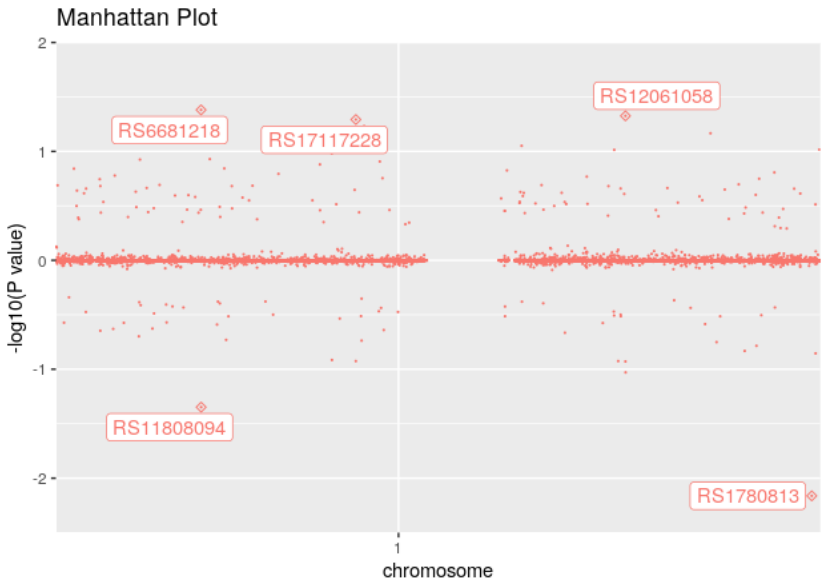

Figure 8: Manhattan Plot of the top 5 SNPs for the lasso method

3.2.5 Results - Frequentist Method

Figure 9 is drawn from the frequentist methods data, highlighting all SNPs of chromosome 1 that have been processed. The annotated SNPs featured are the three lowest p-values including SNP rs17849501. The Bentham et al (Europeans) and the Morris et al (European and Asians) studies found SNP rs17849501 as a genetic risk locus for lupus. This study ranked three of the methods highly while the EBEN produced a zero beta coefficient. The frequentist p-value was extremely low at 1.51E-72.

Table 13: Top ten SNPs ranked for Chromosome 1 and accompanied with their p-value. The associated SNP is highlighted in bold.

Frequentist			
SNP	Position	RANK	P-VALUE
RS17849501	**183542323**	**1**	**1.51E-72**
RS1780813	246444082	2	9.53E-21
RS10911346	183468420	3	5.61E-18
RS17349278	4836929	4	3.33E-17
RS12081621	61943156	5	1.42E-16
RS2298083	183515428	6	1.10E-15
RS789177	183515777	7	1.08E-14
RS3845622	159171603	8	1.94E-13
RS4916334	173333829	9	2.44E-13
RS10798269	173309713	10	3.89E-13

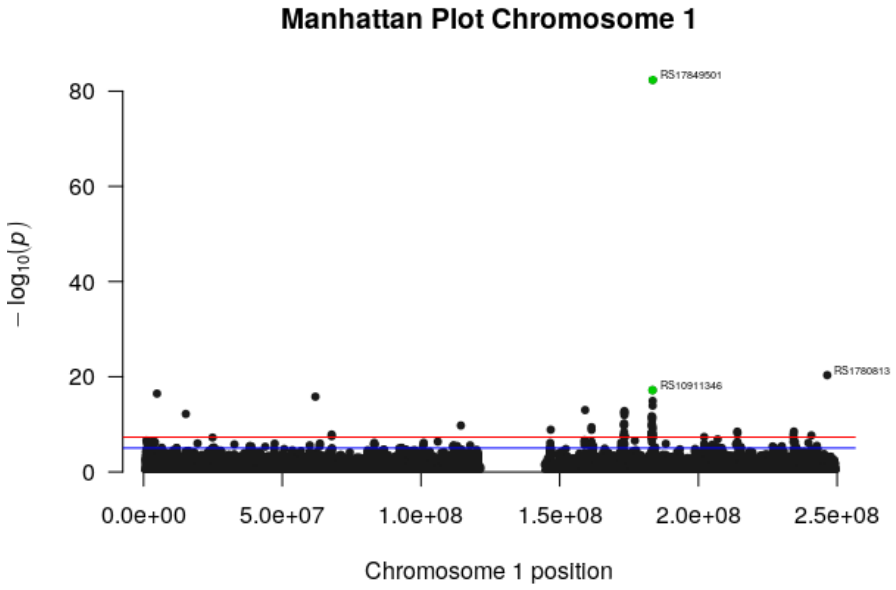

Figure 9: A Manhattan plot with top ranked SNPs annotated from the frequentist methods data

Figure 10 reveals the associated SNP rs17849501 has very little linkage with surrounding SNPs. Contrastively, the two SNPs either side, have high correlation between each other.

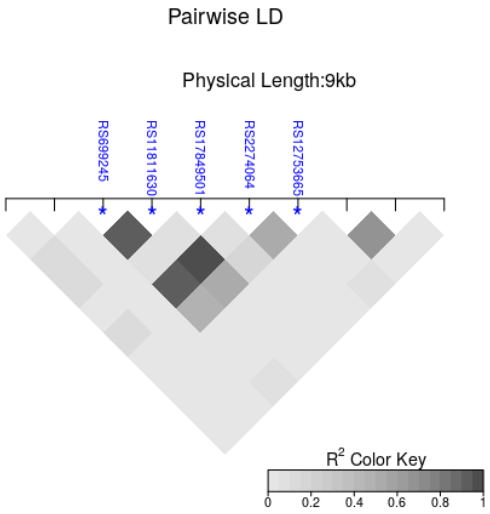

Figure 10: An associated risk loci rs17849501 showing low to zero linkage disequilibrium with surrounding SNPs.

3.2.6 Results - EBEN Method

All SNPs failed to make genome wide significance, although the associated SNP rs2476601 did rank 5th overall.

Table 14: Top ten SNPs ranked for Chromosome 1 and accompanied with their p-value. The associated SNP is highlighted in bold.

EBEN			
SNP	Position	RANK	P-VALUE
RS3845622	159171603	1	1.61E-05
RS1538971	161676394	2	1.44E-04
RS4240539	115603844	3	2.83E-04
RS2476601	**114377568**	**4**	**3.08E-04**
RS1032608	167131321	5	4.35E-04
RS10913245	176705147	6	4.69E-04
RS490800	92592905	7	6.72E-04
RS17494681	181480183	8	8.52E-04
RS901917	92292407	9	9.21E-04
RS10159082	204160450	10	1.01E-03

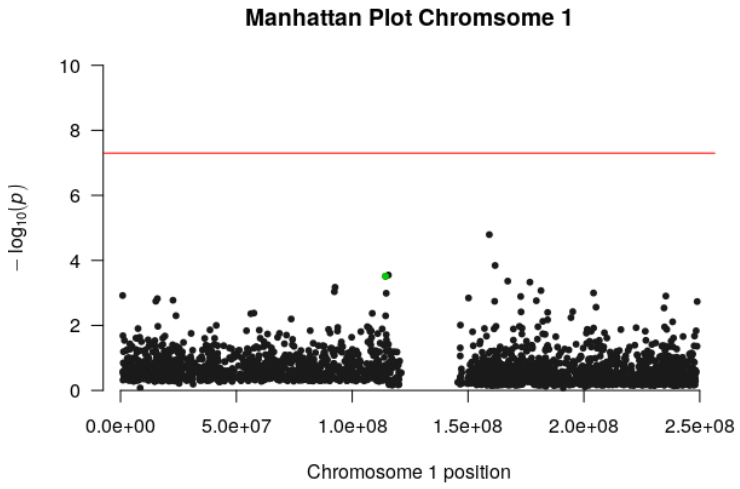

Figure 11: A Manhattan plot with the associated SNP rs2476601 highlighted

3.2.7 Results - Summary

Associated SNP rs10798269, found by Harley et al [41] in 2008, was ranked 10th in this study by the frequentist method, but with a weak result from the EBEN and 2 zero coefficients from the other two methods. The associated SNP rs3024505 fared poorly in this study with the variable selection methods all producing a zero-beta coefficient and an insignificant p-value by the frequentist method. Also, the SNP rs9782955 produced 2 zero coefficients and 2 weak scores in this study. The associated SNP rs525410 produced weak results from the frequentist and EBEN method, while the Spike and the lasso reported zero coefficients. Associated SNP rs10911628 produced three zero coefficients and a poor result from the frequentist method. Associated SNPs featured in the Morris et al study [43] rs34889541 and rs2297550 were not part of this study. In a report by Han et al [102] that involved Asian and European populations, SNP rs1234315 in gene *TNFSF4* was associated with lupus. This SNP was ranked in the top 350 of all four methods and was 22nd by the frequentist method. Bentham et al study found SNP rs704840 to be associated with lupus and was ranked 18th by the frequentist method, while the variable selection methods each had a small non-zero coefficient. Martin et al [100] had previously reported this risk locus 2 years previous, with both studies based on European populations. The associated SNP rs1801274 from Bentham et al study ranked 171st or higher for all methods with the frequentist ranking 28th with a p-value of 1.51E-09 and the EBEN recording 19th rank.

In 2015, Sheng et al [103] produced a study of an association of lupus with SNP rs4916219 to a genome wide level of significance in Chinese population. In this study, the variable selection methods calculated a non-zero beta value and the frequentist method ranked it 47th. SNP rs6662618 had a p-value of 9.76E-06 recorded by the frequentist method.

SNP rs2476601 that was ranked 4th by the EBEN method and ranked 161st and higher by the other methods, has reported links with other disease including autoimmune thyroid disease [104], arthritis [105], and many more including SLE first recorded by Gateva et al [60]. All cases of lupus were in European population studies.

SNP rs35358165 has no association with any disease but has a very high consistency amongst the variable selection methods ranking 12th for the spike, 3rd for the lasso and 20th for EBEN.

Table 15 below shows the SNPs that were highly ranked once the data for chromosome 1 had been analysed using each of the four methods. Each block starting with the spike method produces a set of top ten ranked SNPs, followed by the lasso's top ten that do not feature in the spike's ten and so on. For example, if the methods each found the same 10 SNPs, only 10 SNPs would be listed in the table. If they all found different SNPs, 40 SNPs would be listed. The table therefore allows readers to see how consistent the results are across the four methods. If a zero is featured, this means the method used variable selection to discard this SNP and used a closely located one that is highly correlated with the original SNP. Any figures in bold type are associated SNPs from Bentham et al study. This table therefore provides an overall summary of the results and how they compare with the previous GWAS findings. This is presented in a standardised style throughout the thesis, in an attempt to make the very large number of results to be presented as easy to follow as possible.

For chromosome 1, from 52418 SNPs studied, spike produced 1801 non-zero variables, lasso produced 1701 and EBEN produced 2304.

Table 16 records the coefficients produced from the Spike and the lasso method and shows the p-values from the frequentist and EBEN methods in the exact format as the previous table.

Table 15: Chromosome 1 - Top ten SNPs ranked for each method

Chromosome 1					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS1780813	246444082	1	1	2	274
RS6681218	47959645	2	4	21174	200
RS11808094	47989572	3	58	2281	0
RS12061058	185891053	4	101	904	0
RS17117228	98271048	5	7	24422	223
RS904295	100963041	6	111	81	0
RS17020562	213542706	7	5	101	0
RS6680879	100965683	8	571	3170	0
RS11204930	152151950	9	6	3067	0
RS10911789	185940266	10	596	39501	0
RS17849501	**183542323**	**17**	**2**	**1**	**0**
RS35358165	248900100	12	3	18436	20
RS11811658	6641758	25	8	17614	105
RS2256917	182249873	13	9	3559	77
RS17117203	98265210	16	10	41930	304
RS10911346	183468420	0	0	3	371
RS17349278	4836929	0	0	4	0
RS12081621	61943156	0	0	5	0
RS2298083	183515428	0	0	6	367
RS789177	183515777	0	0	7	0
RS3845622	159171603	0	0	8	1
RS4916334	173333829	0	0	9	2046
RS10798269	173309713	0	0	10	264
RS1538971	161676394	170	199	133	2
RS4240539	115603844	159	74	694	3
RS2476601	**114377568**	**161**	**48**	**25**	**4**
RS1032608	167131321	165	126	13726	5
RS10913245	176705147	232	241	3210	6
RS490800	92592905	176	304	669	7
RS17494681	181480183	196	257	9141	8
RS901917	92292407	172	88	6052	9
RS10159082	204160450	189	259	6065	10

Table 16: Chromosome 1 - Top ten SNPs for each method with their coefficient or p-value. The associated SNP is highlighted in bold.

Chromosome 1				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS1780813	-2.16E+00	-2.16E+00	9.53E-21	7.04E-02
RS6681218	1.38E+00	1.38E+00	2.56E-01	5.25E-02
RS11808094	-1.35E+00	-2.02E-01	5.56E-03	0
RS12061058	1.33E+00	1.42E-01	8.81E-04	0
RS17117228	1.29E+00	1.29E+00	3.21E-01	7.03E-02
RS904295	1.23E+00	1.29E-01	9.74E-07	0
RS17020562	1.17E+00	1.17E+00	3.06E-06	0
RS6680879	1.12E+00	2.73E-02	1.02E-02	0
RS11204930	1.05E+00	1.05E+00	9.52E-03	0
RS10911789	-1.03E+00	-2.61E-02	6.66E-01	0
RS17849501	**-9.25E-01**	**-9.25E-01**	**1.51E-72**	**0**
RS35358165	1.02E+00	1.02E+00	2.05E-01	1.85E-03
RS11811658	8.41E-01	8.41E-01	1.89E-01	2.46E-02
RS2256917	1.01E+00	1.01E+00	1.26E-02	1.82E-02
RS17117203	-9.25E-01	-9.25E-01	7.28E-01	3.30E-01
RS10911346	0	0	5.61E-18	9.95E-02
RS17349278	0	0	3.33E-17	0
RS12081621	0	0	1.42E-16	0
RS2298083	0	0	1.10E-15	3.48E-01
RS789177	0	0	1.08E-14	0
RS3845622	0	0	1.94E-13	1.61E-05
RS4916334	0	0	2.44E-13	5.54E-01
RS10798269	0	0	3.89E-13	3.17E-01
RS1538971	1.06E-01	8.19E-02	7.59E-06	1.44E-04
RS4240539	3.46E-01	1.78E-01	5.19E-04	2.83E-04
RS2476601	**3.31E-01**	**2.17E-01**	**3.20E-10**	**3.08E-04**
RS1032608	1.34E-01	1.21E-01	1.25E-01	4.35E-04
RS10913245	5.72E-02	6.80E-02	1.05E-02	4.69E-04
RS490800	8.97E-02	5.16E-02	4.73E-04	6.72E-04
RS17494681	7.21E-02	6.42E-02	6.39E-02	8.52E-04
RS901917	1.03E-01	1.58E-01	3.13E-02	9.21E-04
RS10159082	7.57E-02	6.31E-02	3.14E-02	1.01E-03

3.3 Chromosome 2

3.3.1 Introduction

Chromosome 2 is the second largest in a human. The first link with lupus (in humans) and chromosome 2 came in 2002 when significant associations on the genes *STAT4* [52] and *PDCD1* [53] were picked up.

3.3.2 Previous GWAS associations in Chromosome 2

In 2014 Armstrong et al [62] linked SNP rs12993006 in gene *BIN1* and rs4544377 in gene *KCNJ3*. A year later Bentham et al reported SNPs rs67040462 in *SPRED2* and rs3768792 in *IKZF2* as risk alleles. The Morris et al study [43] found the gene *LBH* to have two risk alleles (rs7579944 and rs17321999).

Table 17: Timeline of asscociated SNPs with lupus through GWAS 2008-2018.

GWAS 2008-2018					
Year	Chr	Associated SNP	Likely causal gene	Study population	Author
2014	2	rs12993006	*BIN1*	EUR	ARM
2014	2	rs4544377	*KCNJ3*	EUR	ARM
2015	2	rs67040462	*SPRED2*	EUR	BEN
2015	2	rs3768792	*IKZF2*	EUR	BEN
2016	2	rs7579944	*LBH*	EC	MOR
2016	2	rs17321999	*LBH*	EC	MOR

KEY: EUR=European, EC=European and Chinese, EA=European American, BEN=Bentham et al [4], MOR=Morris et al [43], ARM=Armstrong et al [62].

3.3.3 Results - Spike and Slab Method

The top 3 ranked SNPs by the spike and slab method are the same as the lasso method and are located close to each other. Both have no known associations to any disease. In this block of 3 SNPs, the top ranked SNP for the frequentist method SNP rs13019891 and the top ranked SNP for the EBEN SNP rs10165797 are joined by SNP rs9308682. These can be seen in Figures 12 and 14

Table 18: Top ten SNPs ranked for Chromosome 2 and accompanied with their coefficient

Spike and Slab			
SNP	Position	RANK	COEFFICIENT
RS13019891	113829869	1	3.19E+00
RS10165797	113828450	2	-2.91E+00
RS9308682	113828425	3	-2.35E+00
RS2287610	169929059	4	1.53E+00
RS17017386	79866200	5	1.43E+00
RS1906892	163997890	6	1.40E+00
RS10209445	12465306	7	1.38E+00
RS17043443	22376557	8	1.30E+00
RS2192890	36899898	9	1.23E+00
RS6748674	192236620	10	1.19E+00

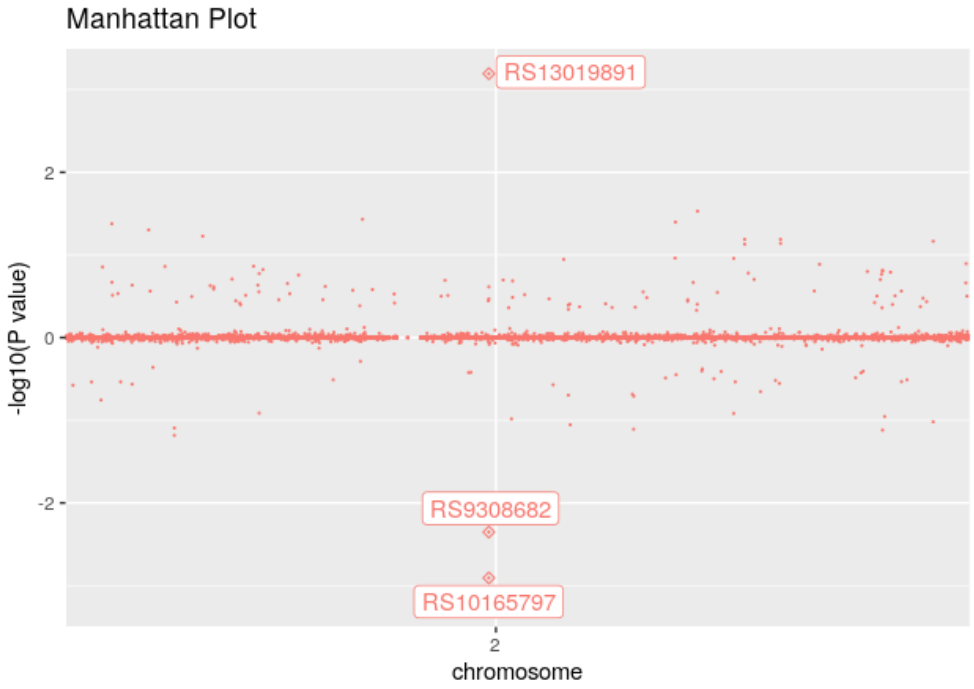

Figure 12: A Manhattan plot with the top three ranked SNPs highlighted for the spike method.

3.3.4 Results - Frequentist Method

Table 19: Top ten SNPs ranked for Chromosome 2 and accompanied with their p-value. The associated SNP is highlighted in bold.

Frequentist			
SNP	Position	RANK	P-VALUE
RS13019891	113829869	1	4.75E-92
RS11889341	**191943742**	**2**	**1.37E-70**
RS7574865	191964633	3	9.25E-67
RS10174238	191973034	4	5.23E-63
RS10168266	191935804	5	1.98E-43
RS2573219	233288667	6	9.78E-39
RS3024886	191900449	7	2.24E-35
RS10931481	191954852	8	1.20E-34
RS3024866	191922841	9	4.70E-24
RS2293765	191520845	10	2.82E-22

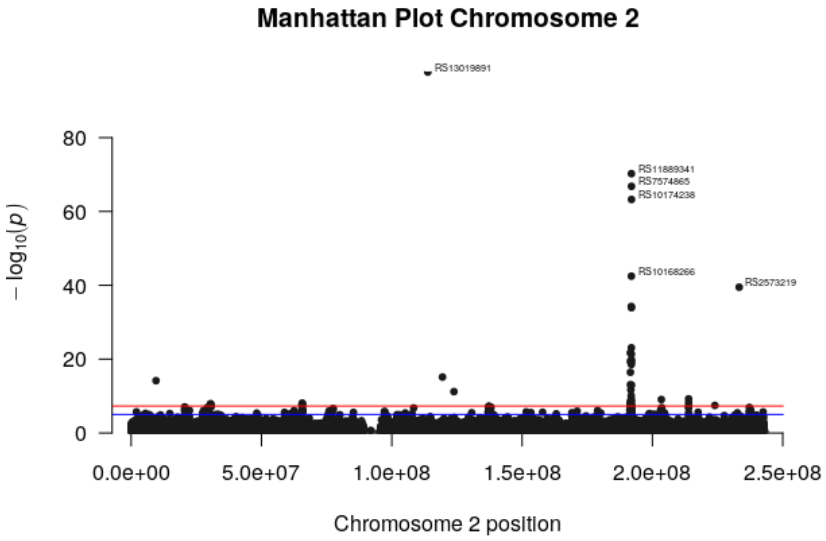

Figure 13: A Manhattan plot with the top 6 ranked SNPs annotated.

3.3.5 Results - Lasso Method

Figure 14 shows beta coefficients. This has been produced from the lasso data and highlights the top 2 extreme valued SNPs rs13019891 and rs10165797.

Table 20: Top ten SNPs ranked for Chromosome 2 and accompanied with their coefficient

Lasso			
SNP	POSITION	RANK	COEFFICIENT
RS13019891	113829869	1	1.88E+00
RS10165797	113828450	2	-1.35E+00
RS9308682	113828425	3	-9.89E-01
RS17043443	22376557	4	6.82E-01
RS4074976	9546695	5	-6.21E-01
RS16831215	135722328	6	-6.16E-01
RS17017386	79866200	7	5.76E-01
RS2287610	169929059	8	5.76E-01
RS3738888	215595164	9	4.93E-01
RS10445824	220236756	10	-4.73E-01

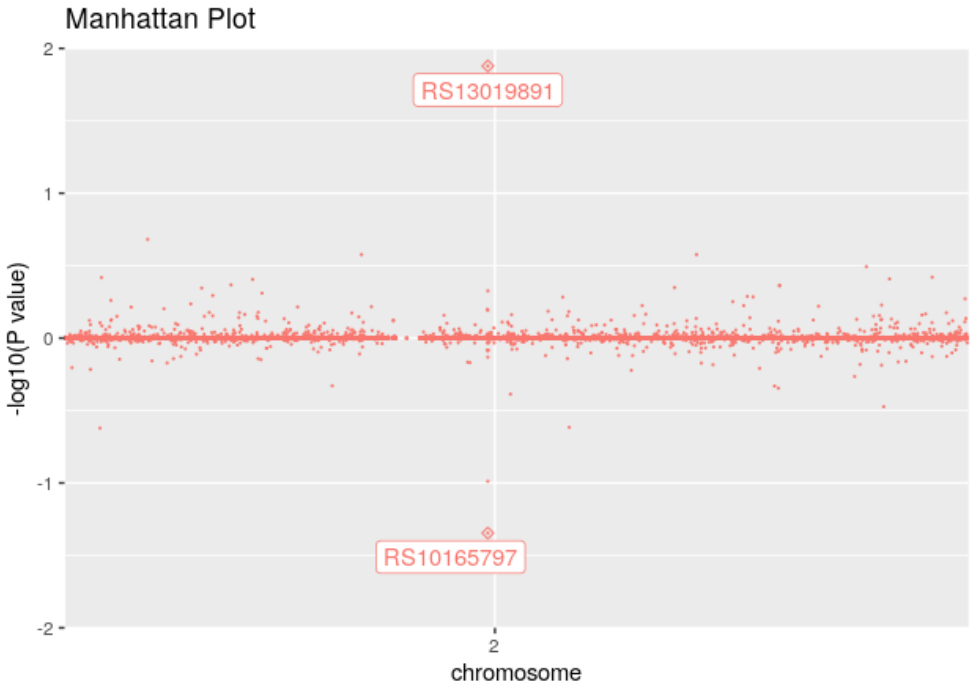

Figure 14: Manhattan plot of the top ranked SNPs for lasso method.

3.3.6 Results - EBEN Method

Table 21: Top ten SNPs ranked for Chromosome 2 and accompanied with their p-value. No associated SNP ranked in the top ten.

EBEN			
SNP	Position	RANK	P-VALUE
RS10165797	113828450	1	4.44E-15
RS2573219	233288667	2	8.15E-14
RS17583054	71342142	3	1.09E-08
RS12373778	47253485	4	3.86E-07
RS2072532	40366301	5	6.31E-07
RS11687809	61003193	6	1.08E-06
RS2339929	24051245	7	1.27E-06
RS6760940	56784574	8	2.34E-06
RS2579500	97201682	9	6.07E-06
RS2309389	98784275	10	7.24E-06

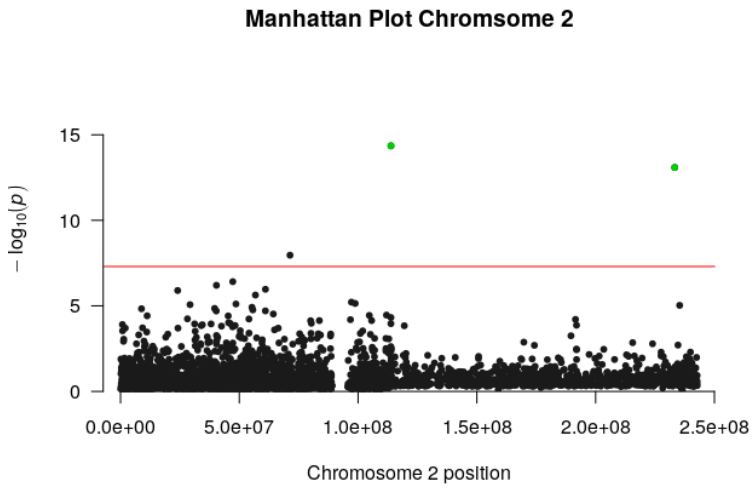

Figure 15: A Manhattan plot with the top 2 ranked SNPs rs10165797 and rs2573219 highlighted.

3.3.7 Results - Summary

For chromosome 2, from 51086 SNPs, spike method produced 1752 non-zero variables, lasso method produced 1556 and the EBEN method produced 3793. SNP rs2573219 was consistent across all four methods with all four being ranked in the top 14. This SNP has no known associations to any disease.

The associated SNPs found by Armstrong in 2014 [62], were rs12993006 and rs4544377. rs12993006 produced 3 zero coefficients with the variable selection methods and a weak p-value from the frequentist method. The SNP rs4544377 was not included in the data used in this study.

Located in the gene *LBH*, SNP rs7579944 was associated with lupus in studies by Morris et al [43] and Langefeld et al [63], both with multi-ethnic cases. The results from this study produced weak beta and p-values. Also in gene *LBH*, Morris et al stated that SNP rs17321999 was a risk locus for lupus. This SNP was not part of this study.

All four methods calculated strong results for rs13023118. It was ranked 111th by spike, 52nd by lasso, frequentist 30th and EBEN 211th.

In a study of Chinese population Yang et al [106] found SNP rs7601754 associated with lupus, as did Martin et al [100] with Europeans as the subjects. This SNP ranked in the top 150 for all methods with the frequentist ranking it 12th.

In Bentham et al study, SNP rs6740462 in the gene *SPRED2* was associated with lupus. This SNP was ranked in the top 200 of the spike, lasso and frequentist methods with a weaker p-value from the EBEN method (ranked 364th). Also, SNP rs2111485 in the gene *IFIH1*, featured in the top 500 ranked SNPs for all methods. This has been previously noted as being a risk locus for vitiligo, psoriasis, inflammatory bowel disease and ulcerative colitis. Bentham et al found another SNP linked to lupus that has been associated with other diseases (rheumatoid arthritis and Sjogren's syndrome). SNP rs11889341 in the gene *STAT4* ranked highly (top 72 or higher) in all four methods in this study including the frequentist method producing a 1.37E-70 p-value and 2nd ranking (as shown in Figure 22).

Also in the Bentham et al study, SNP rs3768792 was found to be a risk locus for SLE. This study found no significant association resulting in low-ranking scores. This SNP is in high linkage disequilibrium with rs10932459, rs13023118, rs10048743 and rs9808132 as shown in Figure 16.

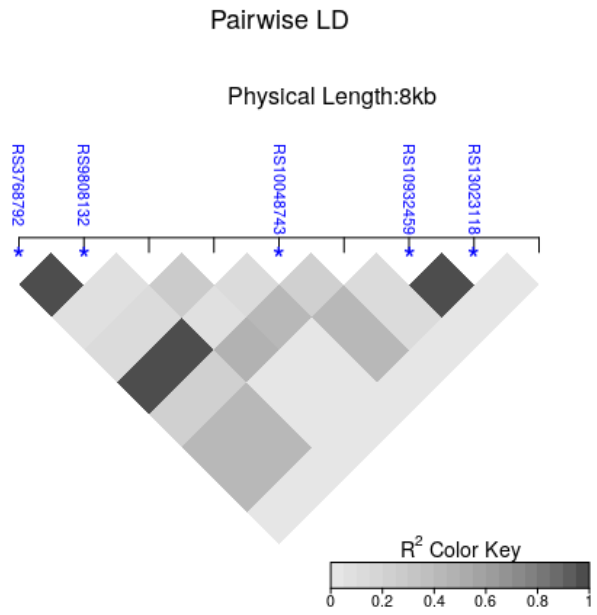

Figure 16: Block of SNPs that have strong LD with the associated risk locus rs3768792

Table 22: Chromosome 2 - Top ten SNPs for each method

Chromosome 2				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS13019891	3.19E+00	1.88E+00	4.75E-92	0
RS10165797	-2.91E+00	-1.35E+00	4.20E-02	4.44E-15
RS9308682	-2.35E+00	-9.89E-01	6.95E-02	4.54E-02
RS2287610	1.53E+00	5.76E-01	2.15E-01	4.63E-02
RS17017386	1.43E+00	5.76E-01	1.44E-02	1.92E-02
RS1906892	1.40E+00	3.49E-01	5.80E-04	0
RS10209445	1.38E+00	2.60E-01	2.56E-03	0
RS17043443	1.30E+00	6.82E-01	7.81E-07	0
RS2192890	1.23E+00	3.45E-01	4.87E-02	1.25E-03
RS6748674	1.19E+00	3.60E-01	2.00E-03	0
RS4074976	-7.55E-01	-6.21E-01	5.37E-15	0
RS16831215	-1.05E+00	-6.16E-01	6.25E-01	0
RS3738888	8.00E-01	4.93E-01	3.69E-03	1.42E-03
RS10445824	-9.54E-01	-4.73E-01	3.04E-02	1.60E-01
RS11889341	**-5.54E-01**	**-3.46E-01**	**1.37E-70**	**1.35E-04**
RS7574865	0	-4.27E-02	9.25E-67	3.68E-03
RS10174238	0	0	5.23E-63	4.36E-01
RS10168266	0	0	1.98E-43	3.45E-01
RS2573219	1.17E+00	4.20E-01	9.78E-39	8.15E-14
RS3024886	0	0	2.24E-35	0
RS10931481	0	0	1.20E-34	1.32E-01
RS3024866	0	0	4.70E-24	0
RS2293765	1.58E+02	-6.93E-02	2.82E-22	6.19E-05
RS17583054	-5.13E-02	-8.07E-02	3.13E-01	1.09E-08
RS12373778	-1.21E-02	-3.34E-02	1.22E-01	3.86E-07
RS2072532	0	0	8.88E-03	6.31E-07
RS11687809	0	0	9.91E-01	1.08E-06
RS2339929	0	2.16E-02	1.10E-02	1.27E-06
RS6760940	2.51E-02	0	2.03E-02	2.34E-06
RS2579500	5.22E-02	1.89E-02	5.36E-04	6.07E-06
RS2309389	-5.02E-02	-2.06E-02	1.36E-02	7.24E-06

Table 23: Chromosome 2 - Top ten SNPs ranked for each method

Chromosome 2					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS13019891	113829869	1	1	1	0
RS10165797	113828450	2	2	7463	1
RS9308682	113828425	3	3	9898	542
RS2287610	169929059	4	8	19153	549
RS17017386	79866200	5	7	4299	327
RS1906892	163997890	6	19	827	0
RS10209445	12465306	7	32	1744	0
RS17043443	22376557	8	4	81	0
RS2192890	36899898	9	21	8073	96
RS6748674	192236620	10	18	1526	0
RS4074976	9546695	41	5	18	0
RS16831215	135722328	19	6	37101	0
RS3738888	215595164	36	9	2104	99
RS10445824	220236756	24	10	6287	1344
RS11889341	**191943742**	**71**	**20**	**2**	**40**
RS7574865	191964633	0	326	3	145
RS10174238	191973034	0	0	4	2897
RS10168266	191935804	0	0	5	2407
RS2573219	233288667	14	11	6	2
RS3024886	191900449	0	0	7	0
RS10931481	191954852	0	0	8	1150
RS3024866	191922841	0	0	9	0
RS2293765	191520845	158	202	10	27
RS17583054	71342142	258	172	23977	3
RS12373778	47253485	1076	422	13693	4
RS2072532	40366301	0	0	3314	5
RS11687809	61003193	0	0	50753	6
RS2339929	24051245	0	618	3740	7
RS6760940	56784574	524	0	5105	8
RS2579500	97201682	252	682	786	9
RS2309389	98784275	272	632	4117	10

3.4 Chromosome 5

3.4.1 Introduction

Chromosome 5 is the fifth largest chromosome in the human body accounting for approximately 181 million base pairs.

3.4.2 Previous GWAS associations in Chromosome 5

An interesting area of chromosome 5 with respect to SLE associations lies in the gene *MIR3142HG*, with reported associations in European populations to SLE positioned just 18kb apart. A report concluded in a European population in 2009 by Gateva [60] that the marker rs2431099 had an association with SLE. In 2011 a study by Chung [61] connected SNP rs2431697 with lupus. Finally, Marquez et al [107] in 2016 in a joint study of SLE and rheumatoid arthritis discovered the risk allele rs4921283. This small group with rs2431697 in the centre has rs2431099 positioned 8kb away upstream with a r^2 values of 0.75 and D' of 0.97 with rs4921283 downstream, a r^2 value of 0.27 and D' of 0.6. In 2015,

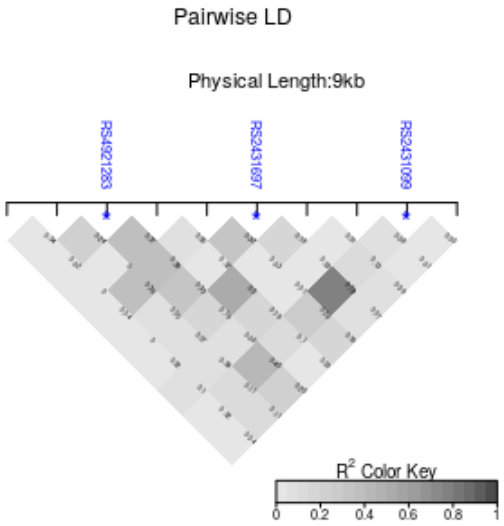

Figure 17: A linkage disequilibrium plot of 3 associated SNPs.

Bentham et al [4] noted rs7726414, located near the genes *TCF7* and *SKP1*, had an association to lupus and this was replicated by Langefeld et al [63] in 2017. Another find came from Julia et al [64] in 2018 with the SNP rs55849330 in the gene *ST8SIA4*. Located 6kb away is SNP rs2548279 with a r^2 value of 0.9 and a D' of 1 to rs55849330. This was also noted as a risk allele in Langefeld et al [63]. Langefeld's study of European, Hispanic and African American populations reported an association with SNP rs461193 in the gene *BC034612* and

SLE. Martin et al [100] in 2013 in a mixed study for systemic sclerosis and systemic lupus erythematosus cases, produced an asociation with SNP rs960709. Associated SNP rs2431099 was observed by all methods with strong results from the frequentist and EBEN methods. This locus was mentioned in the Gateva et al [60] 2009 study of Europeans along with SNP rs7708392. This SNP is in the gene *TNIP1* and was followed up by Alarcon-Riqelme et al [108] and Langefeld et al. This SNP now has links to African American, Hispanic, European and native American populations. SNP rs10036748 has been associated with lupus in East Asian populations [102] followed up by Bentham et al in Europeans and Langefeld et al in Hispanic, Afro Caribbean, and African Americans. This SNP has also been associated with Psoriasis [109]. In this study it ranked top ten in the frequentist and EBEN methods. Both spike and the lasso ranked it in their top 200.

Table 24: Timeline of asscociated SNPs with lupus through GWAS 2008-2018.

GWAS 2008-2018					
Year	Chr	Associated SNP	Likely causal gene	Study Population	Author
2009	5	rs7708392	*TNIP1*	EA	GAT
2011	5	rs2431697	*PTTG1*	EUR	CHU
2015	5	rs7726414	*TCF7,SKP1*	EUR	BEN
2018	5	rs55849330	*ST85IA4*	EUR	JUL

KEY: EUR=European, EC=European and Chinese, EA=European American, BEN=Bentham et al [4], JUL=Julia et al [64], GAT=Gateva et al [60] , CHU=Chung et al [61].

3.4.3 Results - Spike and Slab Method

From the top 16 SNPs ranked by spike, 11 of the lasso are the same. For chromosome 5, the results are very similar for the two methods. They also share rs1078324 and rs11739489 in their top 3 as shown in Figures 18 and 19.

Table 25: Top ten SNPs ranked for Chromosome 5 and accompanied with their coefficient

Spike and Slab			
SNP	POSITION	RANK	COEFFICIENT
RS1078324	149202268	1	-3.20E+00
RS741580	149201670	2	2.34E+00
RS11739489	76325361	3	-2.08E+00
RS16875609	5470026	4	1.78E+00
RS16875597	5460434	5	1.60E+00
RS17087649	96971199	6	1.31E+00
RS2003604	149205940	7	1.30E+00
RS6872933	107336419	8	-1.28E+00
RS1107233	60020520	9	1.27E+00
RS7733069	60057836	10	1.21E+00

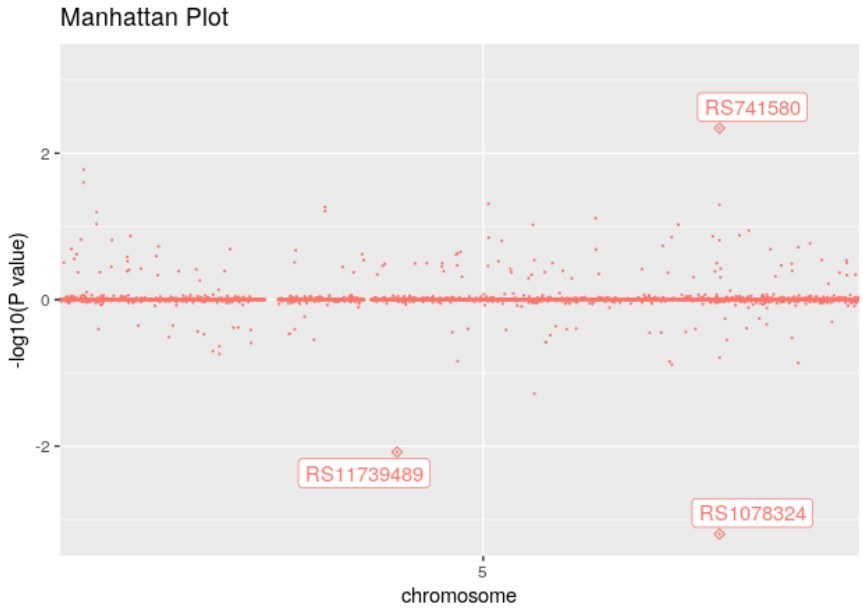

Figure 18: A Manhattan plot with the top SNPs highlighted for spike method

3.4.4 Results - Lasso Method

Another demonstration that the spike and lasso methods are similar sharing 6 markers that feature in both of their lists of top 8 SNPs, of which neither contain previously associated loci.

Table 26: Top ten SNPs ranked for Chromosome 5 and accompanied with their coefficient

Lasso			
SNP	POSITION	RANK	COEFFICIENT
RS11739489	76325361	1	-1.54E+00
RS1078324	149202268	2	-1.53E+00
RS6872933	107336419	3	-1.07E+00
RS741580	149201670	4	9.27E-01
RS1394603	155811990	5	6.94E-01
RS16875609	5470026	6	6.62E-01
RS16875597	5460434	7	6.53E-01
RS6883894	107037375	8	6.35E-01
RS7714120	137919850	9	-5.56E-01
RS1025488	139892331	10	5.37E-01

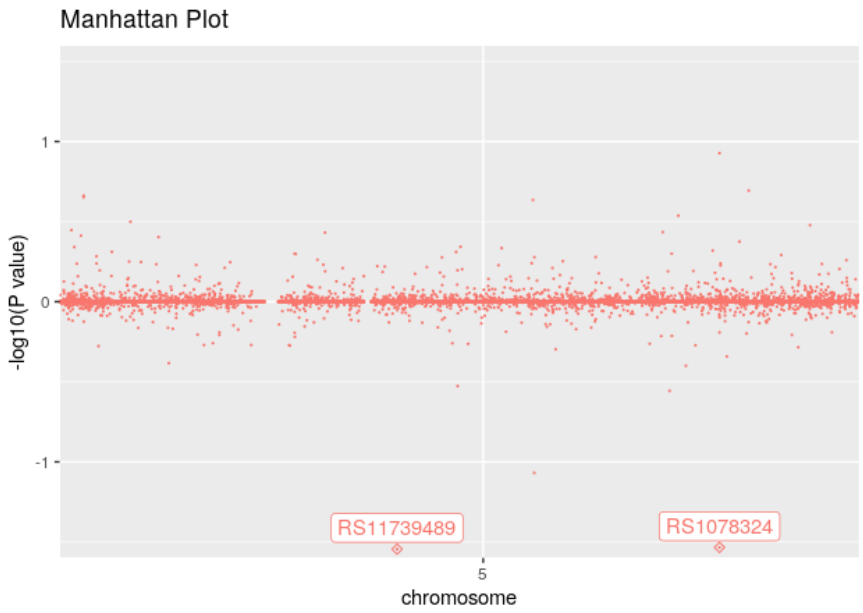

Figure 19: Manhattan plot of the top SNPs for lasso method

3.4.5 Results - Frequentist Method

Out of the top ten SNPs ranked by the frequentist method, six are positioned in a tightly grouped block, 10kb in length. It can clearly be seen from Figure 20 the medium-high linkage disequilibrium amongst all six.

Table 27: Top ten SNPs ranked for Chromosome 5 and accompanied with their coefficient

Frequentist			
SNP	POSITION	RANK	COEFFICIENT
RS1078324	149202268	1	1.76E-22
RS10036748	**150458146**	**2**	**3.30E-22**
RS960709	150461049	3	5.43E-22
RS1422673	150438988	4	9.71E-19
RS7726414	**133431834**	**5**	**4.66E-17**
RS3792785	150451650	6	2.58E-15
RS3792783	150455732	7	2.56E-14
RS2431697	**159879978**	**8**	**3.82E-14**
RS2233287	150440097	9	5.35E-14
RS2431099	159886620	10	8.95E-14

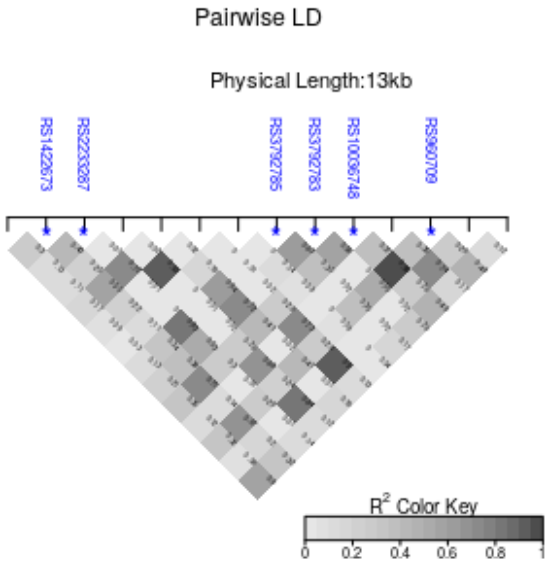

Figure 20: A LDHeatmap showing 6 of the top 10 SNPs ranked by the frequentist method

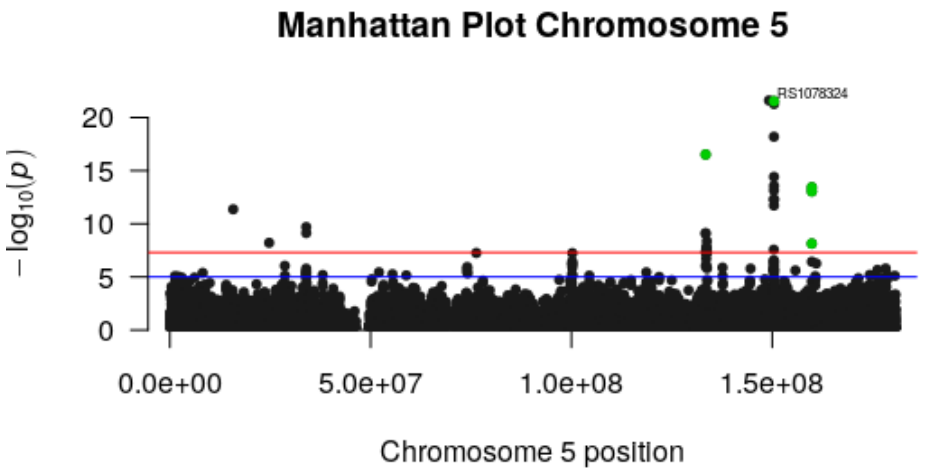

Figure 21: A Manhattan plot with the lowest p-value SNP rs1078324 is annotated with four known hits highlighted in green.

3.4.6 Results - EBEN Method

Two of the previous associated markers feature in the top ten SNPs but none achieved p-values that were statistically significant.

Table 28: Top ten SNPs ranked for Chromosome 5 and accompanied with their coefficient

EBEN			
SNP	POSITION	RANK	COEFFICIENT
RS2431697	**159879978**	**1**	**3.83E-03**
RS10037643	60565864	2	5.95E-03
RS6872933	107336419	3	7.02E-03
RS2431099	159886620	4	7.04E-03
RS10036748	**150458146**	**5**	**7.69E-03**
RS7737958	113067189	6	2.12E-02
RS447817	155651950	7	3.06E-02
RS11241783	124358513	8	3.41E-02
RS4700181	56574452	9	3.66E-02
RS4958296	151612227	10	3.70E-02

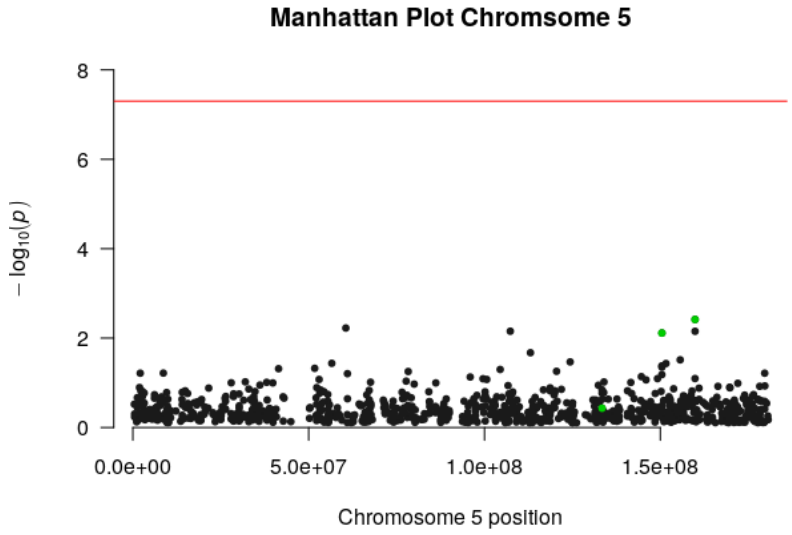

Figure 22: A Manhattan plot with associated SNPs highlighted

3.4.7 Results - Summary

In this study SNP rs2431697 was successfully found by all four methods and with the EBEN method calculating it to be the top ranked SNP. All methods found associated SNP rs7726414 with the spike, lasso and the frequentist methods ranked in the top 100. Langefeld et al [63] study of European, Hispanic and African American populations reported an association with SNP rs461193 in the gene *BC034612* and SLE. All four methods found this SNP but with weak scores. This study also found SNP rs4921283 with all four methods, including the frequentist and EBEN methods in their top 25 ranked SNPs.

The associated SNP rs960709, was 3rd in the frequentist method ranking and was also picked out by EBEN.

The marker rs2431099 was observed by all methods with strong results from the frequentist and EBEN methods. This locus was mentioned in the Gateva et al study of Europeans along with SNP rs7708392 [60]. This SNP is in the gene *TNIP1* and was followed up by Alarcon-Riqelme et al [108] and Langefeld et al [63]. This SNP now has links to African American, Hispanic, European and native American populations.

Although SNP rs6872933 was found in all four methods and ranked in the top ten by three of them it has no previous associations with lupus.

For chromosome 5, from 37924 SNPs, the spike method produced 1368 non-zero variables, the lasso method produced 2605 and the EBEN method produced 850.

Table 29: Chromosome 5 - Top ten SNPs for each method

Chromosome 5				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p- value	EBEN p-value
RS1078324	-3.20E+00	-1.53E+00	1.76E-22	8.07E-02
RS741580	2.34E+00	9.27E-01	9.27E-01	0
RS11739489	-2.08E+00	-1.54E+00	5.40E-08	0
RS16875609	1.78E+00	6.62E-01	4.60E-04	0
RS16875597	1.60E+00	6.53E-01	7.47E-01	4.91E-01
RS17087649	1.31E+00	1.42E-01	2.19E-05	2.27E-01
RS2003604	1.30E+00	3.54E-02	4.70E-01	0.00E+00
RS6872933	-1.28E+00	-1.07E+00	2.00E-03	7.02E-03
RS1107233	1.27E+00	4.32E-01	5.05E-01	0
RS7733069	1.21E+00	1.90E-01	1.88E-01	0
RS1394603	9.45E-01	6.94E-01	5.24E-02	2.54E-01
RS6883894	1.02E+00	6.35E-01	1.31E-01	2.59E-01
RS7714120	-8.45E-01	-5.56E-01	5.00E-03	0
RS1025488	1.02E+00	5.37E-01	6.30E-02	0
RS10036748	**-2.60E-01**	**-1.22E-01**	**3.30E-22**	**7.69E-03**
RS960709	0	0	5.43E-22	3.30E-01
RS1422673	0	-5.87E-03	9.71E-19	6.57E-02
RS7726414	**-4.52E-01**	**-2.63E-01**	**4.66E-17**	**3.70E-01**
RS3792785	0	-8.65E-02	2.58E-15	3.35E-01
RS3792783	0	0	2.56E-14	3.33E-01
RS2431697	**-1.14E-01**	**-1.30E-01**	**3.82E-14**	**3.83E-03**
RS2233287	0	0	5.35E-14	0
RS2431099	3.97E-02	1.55E-02	8.95E-14	7.04E-03
RS10037643	-7.18E-02	-6.07E-02	1.04E-04	5.95E-03
RS7737958	8.85E-02	9.23E-02	1.46E-03	2.12E-02
RS447817	4.10E-02	2.83E-02	1.18E-03	3.06E-02
RS11241783	-1.09E-01	-9.56E-02	2.43E-04	3.41E-02
RS4700181	4.62E-02	4.62E-02	8.73E-03	3.66E-02
RS4958296	1.20E-01	1.29E-01	1.24E-02	3.70E-02

Table 30: Chromosome 5 - Top ten SNPs ranked for each method

Chromosome 5					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS1078324	149202268	1	2	1	25
RS741580	149201670	2	4	35772	0
RS11739489	76325361	3	1	28	0
RS16875609	5470026	4	6	362	0
RS16875597	5460434	5	7	30630	476
RS17087649	96971199	6	132	100	127
RS2003604	149205940	7	771	22268	0
RS6872933	107336419	8	3	814	3
RS1107233	60020520	9	16	23373	0
RS7733069	60057836	10	118	12304	0
RS1394603	155811990	16	5	5663	167
RS6883894	107037375	15	8	9865	176
RS7714120	137919850	24	9	1397	0
RS1025488	139892331	14	10	6336	0
RS10036748	**150458146**	**139**	**167**	**2**	**5**
RS960709	150461049	0	0	3	255
RS1422673	150438988	0	2078	4	21
RS7726414	**133431834**	**82**	**46**	**5**	**306**
RS3792785	150451650	0	250	6	268
RS3792783	150455732	0	0	7	262
RS2431697	**159879978**	**145**	**149**	**8**	**1**
RS2233287	150440097	0	0	9	0
RS2431099	159886620	284	761	10	4
RS10037643	60565864	166	388	186	2
RS7737958	113067189	151	237	681	6
RS447817	155651950	271	971	603	7
RS11241783	124358513	150	230	261	8
RS4700181	56574452	137	573	1948	9
RS4958296	151612227	93	127	2393	10

The associated SNPs rs7708392 found by Gateva et al, Langefeld et al rs2548279 and rs55849330 found by Julia et al [64] were not part of this study.

3.5 Chromosome 6

3.5.1 Introduction

Inside chromosome 6 lies the dense genetic system, the Major Histocompatiblity Complex (MHC). As discussed in subsection 1.4.3., this area has the highest linkage disequilibrium in the human genome. Many studies were undertaken pre-GWAS in small numbers of subjects to look into links between SLE and the gene *HLA-DRB1*. Dong et al [110] in 1993 studied Japanese patients, Brennan et al [57] in 1997 undertook a joint study into Rheumatoid Arthritis and SLE, and Reveille et al [111] in 2004 used subjects from multi-ethnicities.

3.5.2 Previous GWAS associations in Chromosome 6

Graham et al in 2008 [59] produced the first GWAS linking rs5029939 in the gene *TNFAIP3* with lupus. Many more followed in 2009 from the study by Gateva et al [60] of European cases showing SNPs rs6568431 (*PRDM1*), rs11755393 (*UHRFIBP1*) and rs9271366 (*HLA-DRB1*) having statistical significance. Han et al [102] and Yang et al [112] reported *HLA-DRB1* in East Asian cases in 2009 and 2012 respectively. Figure 23 clearly shows the MHC's dense region of complex highly linked SNPs.

In a Gateva et al study with European ancestries, an associated SNP rs6568431 located in the gene *PRDM1* was found with follow up studies including Morris et al and Bentham et al agreeing with the findings. In Bentham et al [4] study, it was reported that SNP rs1270942 has an association with lupus. This SNP has previously been associated with autoimmune thyroid disease and type I diabetes. In a study by Armstrong et al [62] SNP rs9275572 (in the HLA region) reported an association with lupus, and has associations with hepatitis and alopecia too. In studies by Chen et al and Langefeld et al they both found SNP rs2327832 in the gene *OLIG3* to have an association with lupus. In 2011, Chung et al [61] found the SNP rs1150754 to be a risk allele. In a study by Morris et al [43] with Chinese and European populations, they found SNP rs597325 in the gene *BACH2* to have an association, also noted was SNP rs17603856 as a risk allele.

Table 31: Timeline of asscociated SNPs with lupus through GWAS 2008-2018.

GWAS 2008-2018					
Year	Chr	Associated SNP	Likely causal gene	Study population	Author
2008	6	rs5029939	*TNFAIP3*	EUR	GRA
2009	6	rs6568431	*PRDM1*	EA	GAT
2009	6	rs11755393	*UHRF1BP1*	EA	GAT
2011	6	rs1150754	*TNXB*	EUR	CHU
2016	6	rs17603856	*ATXN1*	EC	MOR
2016	6	rs597325	*BACH2*	EC	MOR
2017	6	rs10498722	*LRRC16A*	EA	LAN
2017	6	rs4712969	*SLC17A4*	EA	LAN
2017	6	rs2327832	*OLIG3-LOC100130476*	EA	LAN

KEY: EUR=European, EC=European and Chinese, EA=European American, MOR=Morris et al [43], LAN=Langefeld et al [63], GAT=Gateva et al [60] , CHU=Chung et al [61], GRA=Graham et al [59].

3.5.3 Results - Spike and Slab Method

A very similar result from the spike and the lasso methods was observed again. Of the top 15 SNPs by both methods, 11 of the same SNPs appear.

Table 32: Top ten SNPs ranked for Chromosome 6 and accompanied with their coefficient

Spike and Slab			
SNP	POSITION	RANK	COEFFICIENT
RS9969061	71862393	1	-2.73E+00
RS2517490	31038756	2	1.91E+00
RS505997	32121932	3	-1.78E+00
RS9396560	15136789	4	-1.73E+00
RS2517491	31038338	5	1.54E+00
RS16895550	164339125	6	-1.47E+00
RS2849013	32132590	7	-1.44E+00
RS7747637	15172268	8	1.40E+00
RS11962557	20407282	9	1.24E+00
RS9501430	30256841	10	1.21E+00

3.5.4 Results - Lasso Method

Table 33: Top ten SNPs ranked for Chromosome 6 and accompanied with their coefficient

Lasso			
SNP	POSITION	RANK	COEFFICIENT
RS9969061	71862393	1	-1.88E+00
RS16895550	164339125	2	-1.04E+00
RS11962557	20407282	3	7.70E-01
RS710090	106644953	4	6.87E-01
RS9501430	30256841	5	6.11E-01
RS9396560	15136789	6	-5.84E-01
RS17064525	104460722	7	-5.61E-01
RS6935887	143047226	8	4.87E-01
RS17073598	144622457	9	4.81E-01
RS11962528	68611523	10	4.73E-01

3.5.5 Results - Frequentist Method

In complete contrast to the previous methods results, the top 1500 SNPs ranked by the frequentist method produced sparse non-zero coefficients with the other three methods. This is likely caused by the dense linkage of the MHC, which is clearly visible on the Manhattan plot in Figure 23. The top 1593 SNPs by the frequentist method were below the statistically significant level of p-value of 5E-08.

Table 34: Top ten SNPs ranked for Chromosome 6 and accompanied with their coefficient

Frequentist			
SNP	POSITION	RANK	COEFFICIENT
RS2854275	32628428	1	7.86E-94
RS9273327	32623223	2	5.04E-92
RS2187668	32605884	3	1.39E-91
RS1270942	**31918860**	**4**	**6.31E-81**
RS519417	31878433	5	2.18E-80
RS558702	31870326	6	1.33E-79
RS497309	31892484	7	1.45E-79
RS3117574	31725230	8	1.52E-79
RS1150753	32059867	9	1.93E-79
RS3101017	31733466	10	3.13E-79

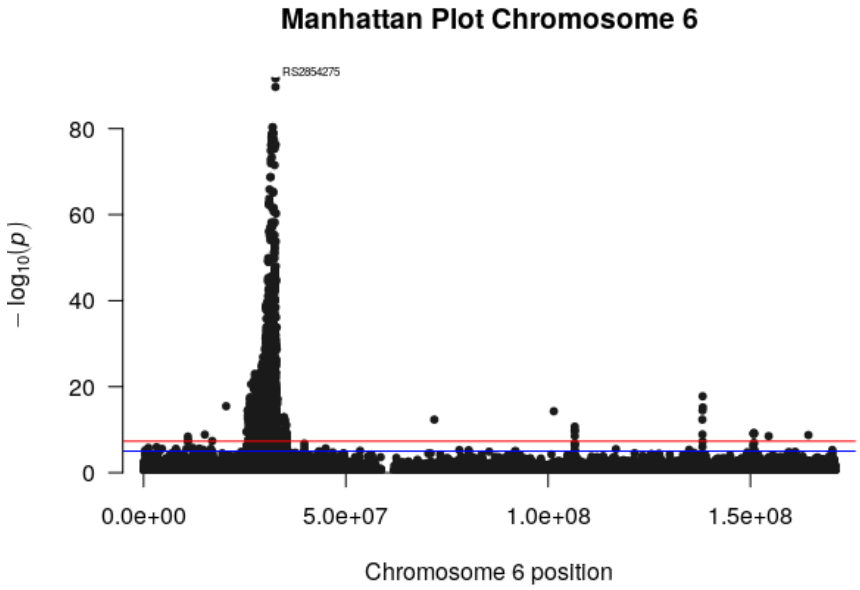

Figure 23: A Manhattan plot exhibiting the densely packed SNPs of the Major Histocompatibility Complex annotated with rs2854275 from the frequentist method's data.

3.5.6 Results - EBEN Method

Table 35: Top ten SNPs ranked for Chromosome 6 and accompanied with their coefficient

EBEN			
SNP	POSITION	RANK	COEFFICIENT
RS2327832	137973068	1	5.79E-06
RS7452689	104016969	2	1.49E-05
RS3104406	32682443	3	2.15E-05
RS6568431	**106588806**	**4**	**2.27E-05**
RS17064525	104460722	5	4.33E-05
RS9375268	123898505	6	9.04E-05
RS6919638	3756043	7	9.21E-05
RS670369	138147048	8	1.31E-04
RS2517491	31038338	9	1.44E-04
RS9402743	136001034	10	2.62E-04

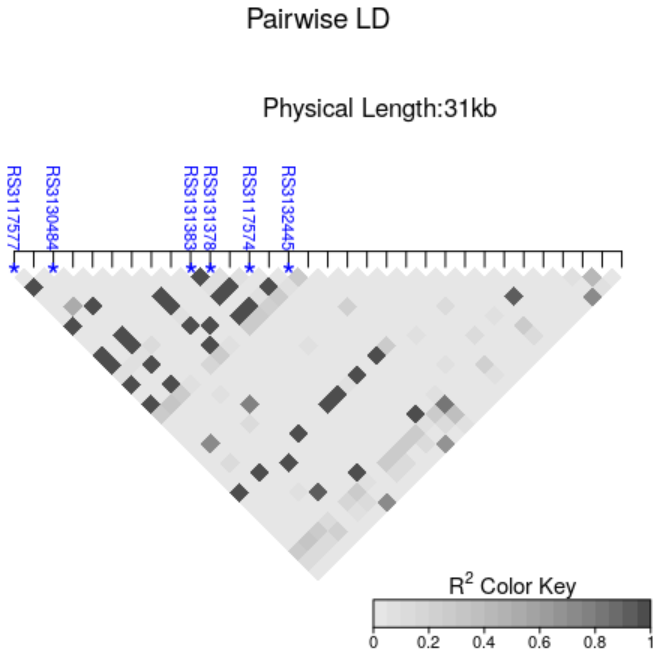

Figure 24: The block showing the 6 SNPs all ranked between 8th-21st by the frequentist method in near perfect disequilibrium

3.5.7 Results - Summary

For chromosome 6, from 42993 SNPs, spike method produced 1255 non-zero variables, lasso method produced 2309 and EBEN produced 488. The associated SNP rs6568431 located in the gene *PRDM1* was found by all four methods including a ranking of 4th by the EBEN method. The risk allele rs9271366 was recorded with zero coefficients by the variable selection methods and achieved a weak result from the frequentist method. Another associated SNP rs11755393, again produced zero coefficients for the variable selection methods, a ranking of only 1420th for the frequentist although it did record a statistically significant p-value of 3.03E-09. Associated SNP rs6568431 was ranked 4th by EBEN method, 144th (lasso), 167th (spike), and 1112th by frequentist method with a p-value of 2.06E-1.

The associated SNP rs6932056 produced 2 zero beta coefficients in the variable selection methods. Positioned 3 SNPs away is rs2230926. This SNP was ranked 65th by the spike and slab and 55th by the lasso. It also produced better results for the frequentist and EBEN methods than rs6932056 did. The pair have LD calculations of $R^2 = 0.839$ and D' = 0.972. Only the frequentist method found the associated SNP rs1270942 resulting in a 4th rank and a p-value of 6.31E-81. The other three methods all made this a zero beta co-

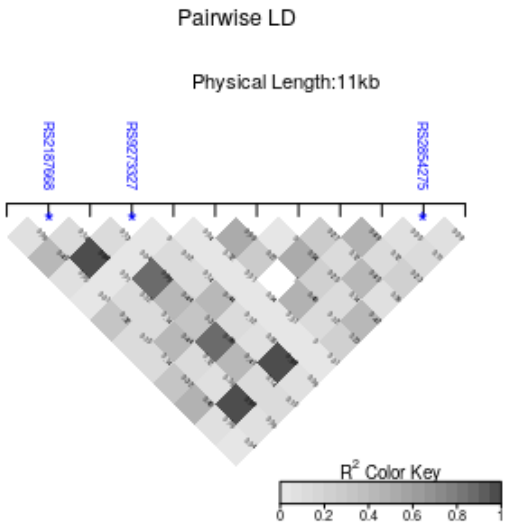

Figure 25: The block shows the top 3 SNPs all ranked by the frequentist method in near perfect disequilibrium

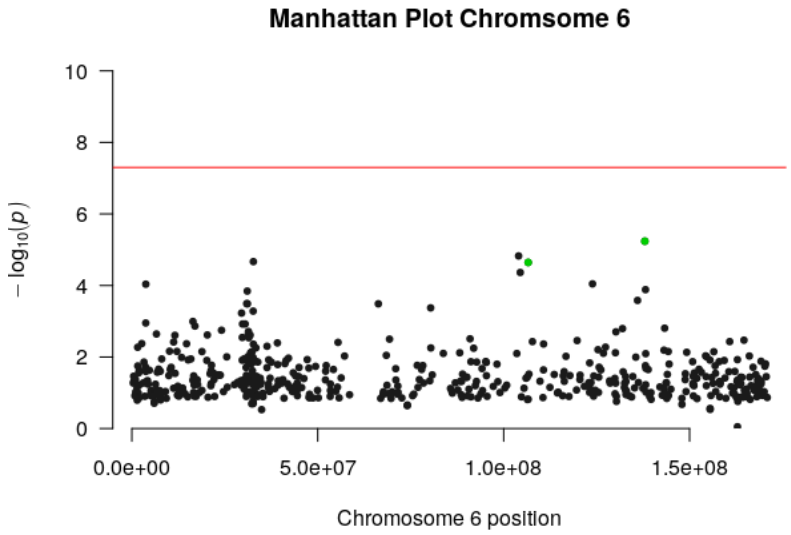

Figure 26: A Manhattan plot for chromosome 6 highlighting rs6568431 and rs2327832 using the EBEN data.

efficient. Spike and lasso methods produced the nearby SNP rs2072634 to be ranked 39th and 30th respectively although there is no LD between these two markers. For the associated SNP rs9275572, this study produced a p-value of 4.86e-48 ranked 79th by the frequentist method but were given zero coefficients by the others. SNP rs2187668 (ranked 3rd by frequentist method) is a tag SNP for the *HLA-DRB1*0301* allele and was reported to have the highest risk for developing SLE in a study of UK family based association study by Fernando et al [48]. The same study reported SNP rs419788 (frequentist method p-value of 1.31e-34) and a SNP found in the gene *SKIV2L* also had an association. All four methods found rs2327832 with the EBEN producing a top hit with a coefficient of 5.79E-06. For the associated SNP rs1150754 the frequentist method produced a p-value of 6.07E-65 and ranked this 37th. The variable selection methods produced two zero coefficients and a weak beta score. The variable selection methods all failed to report a hit for the associated SNP rs597325. The SNP rs10807150 in gene *DEF6* from a study by Sun et al [113] in an Asian population failed to produce a hit in this study.

Table 36: Chromosome 6 - Top ten SNPs ranked for each method

Chromosome 6					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS9969061	71862393	1	1	945	0
RS2517490	31038756	2	813	41124	0
RS505997	32121932	3	11	1522	0
RS9396560	15136789	4	6	1386	86
RS2517491	31038338	5	26	38469	9
RS16895550	164339125	6	2	1438	0
RS2849013	32132590	7	90	7261	0
RS7747637	15172268	8	24	30504	0
RS11962557	20407282	9	3	695	0
RS9501430	30256841	10	5	16439	80
RS710090	106644953	12	4	2630	0
RS17064525	104460722	15	7	10304	5
RS6935887	143047226	31	8	4698	320
RS17073598	144622457	23	9	28705	0
RS11962528	68611523	14	10	36866	286
RS2854275	32628428	0	0	1	479
RS9273327	32623223	0	47	2	14
RS2187668	32605884	0	0	3	474
RS1270942	**31918860**	**0**	**0**	**4**	**0**
RS519417	31878433	0	0	5	0
RS558702	31870326	0	0	6	0
RS497309	31892484	0	0	7	0
RS3117574	31725230	0	0	8	0
RS1150753	32059867	0	0	9	0
RS3101017	31733466	0	0	10	0
RS2327832	137973068	168	115	3222	1
RS7452689	104016969	171	117	4185	2
RS3104406	32682443	0	284	362	3
RS6568431	**106588806**	**167**	**145**	**1112**	**4**
RS9375268	123898505	194	159	2958	6
RS6919638	3756043	210	155	2284	7
RS670369	138147048	373	201	1372	8
RS9402743	136001034	306	1041	2227	10

Table 37: Chromosome 6 - Top ten SNPs for each method

Chromosome 6				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS9969061	-2.73E+00	-1.88E+00	4.39E-13	0
RS2517490	1.91E+00	3.02E-02	9.37E-01	0
RS505997	-1.78E+00	-4.61E-01	1.32E-08	0
RS9396560	-1.73E+00	-5.84E-01	1.67E-09	1.21E-02
RS2517491	1.54E+00	3.62E-01	8.47E-01	1.44E-04
RS16895550	-1.47E+00	-1.04E+00	3.83E-09	0
RS2849013	-1.44E+00	-1.72E-01	2.68E-02	0
RS7747637	1.40E+00	3.73E-01	5.86E-01	0
RS11962557	1.24E+00	7.70E-01	3.20E-16	0
RS9501430	1.21E+00	6.11E-01	1.91E-01	1.12E-02
RS710090	1.01E+00	6.87E-01	1.10E-04	0
RS17064525	-9.32E-01	-5.61E-01	6.79E-02	4.33E-05
RS6935887	7.20E-01	4.87E-01	5.55E-03	7.49E-02
RS17073598	8.05E-01	4.81E-01	5.28E-01	0
RS11962528	9.33E-01	4.73E-01	7.94E-01	6.19E-02
RS2854275	0	0	7.86E-94	2.01E-01
RS9273327	0	-2.58E-01	5.04E-92	5.22E-04
RS2187668	0	0	1.39E-91	1.72E-01
RS1270942	**0**	**0**	**6.31E-81**	**0**
RS519417	0	0	2.18E-80	0
RS558702	0	0	1.33E-79	0
RS497309	0	0	1.45E-79	0
RS3117574	0	0	1.52E-79	0
RS1150753	0	0	1.93E-79	0
RS3101017	0	0	3.13E-79	0
RS2327832	-2.01E-01	-1.48E-01	5.83E-04	5.79E-06
RS7452689	-1.01E-01	-1.47E-01	2.93E-03	1.49E-05
RS3104406	0	-7.91E-02	1.23E-24	2.15E-05
RS6568431	**2.27E-01**	**1.28E-01**	**2.06E-11**	**2.27E-05**
RS9375268	6.48E-02	1.20E-01	2.88E-04	9.04E-05
RS6919638	4.92E-01	1.22E-01	2.61E-05	9.21E-05
RS670369	3.32E-02	1.01E-01	1.39E-09	1.31E-04
RS9402743	-4.08E-02	-5.06E-03	2.12E-05	2.62E-04

The associated SNPs rs5029939 [59], rs10498722 and rs4712969 [63], rs2230926 and rs17603856 [43] in Europeans and Lessard et al [114] in East Asians were not part of this study.

3.6 Chromosome 16

3.6.1 Introduction

Chromosome 16 has approximately 90 million base pairs. Pre-GWAS links with SLE through Hispanic populations were reported in 2004 [115].

3.6.2 Previous GWAS associations in Chromosome 16

Early studies in 2008 by Harley et al and Hom et al found associated SNPs in chromosome 16, rs11574637 located in the gene *ITGAX* and the SNP rs9888739 in the gene *ITGAM*. Bentham et al report made an association with lupus and SNP rs9652601 as did Morris et al in 2016 with SNP rs1170426. SNP rs12599402 in the gene *CLEC16A* was found to have an association in a Chinese population in a study by Zhang et al [65].

Table 38: Timeline of associated SNPs with lupus through GWAS 2008-2018.

GWAS 2008-2018					
Year	Chr	Associated SNP	Likely causal gene	Study population	Author
2008	16	rs11574637	*ITGAX*	EA	HOM
2008	16	rs9888739	*ITGAM*	EUR	HAR
2015	16	rs9652601	*CIITA,SOCS1*	EUR	BEN
2016	16	rs1170426	*ZFP90*	EC	MOR

KEY: EUR=European, EC=European and Chinese, EA=European American, BEN=Bentham et al [4], MOR=Morris et al [43], HAR=Harley et al [41], HOM=Hom et al [7].

3.6.3 Results - Spike and Slab Method

Although SNP rs28707189 was ranked top by spike it has no previous associations with any disease.

Table 39: Top ten SNPs ranked for Chromosome 16 and accompanied with their coefficient

Spike and Slab			
SNP	POSITION	RANK	COEFFICIENT
RS28707189	2566752	1	2.64E+00
RS8059824	67963284	2	1.50E+00
RS16957597	67946356	3	1.34E+00
RS13335252	67808212	4	1.31E+00
RS8058306	67440289	5	1.23E+00
RS9930906	2534525	6	1.18E+00
RS7187296	3036551	7	-1.01E+00
RS16965197	62855866	8	9.87E-01
RS1486445	51857372	9	9.84E-01
RS2734743	4937909	10	9.68E-01

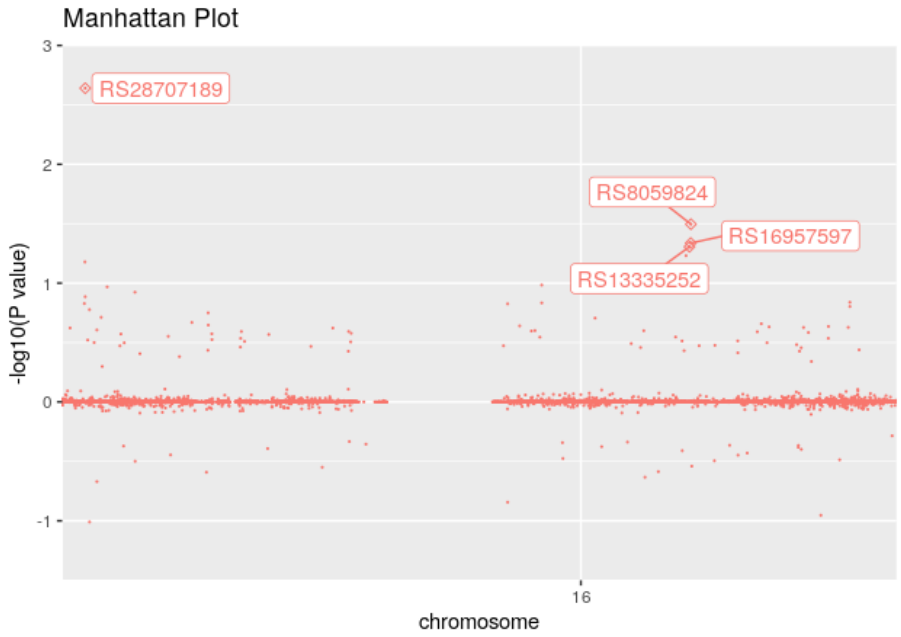

Figure 27: A Manhattan plot with the top SNPs annotated for spike method

3.6.4 Results - Lasso Method

Although SNP rs28707189 was ranked top by lasso it has no previous associations with any disease.

Table 40: Top ten SNPs ranked for Chromosome 16 and accompanied with their coefficient

Lasso			
SNP	POSITION	RANK	COEFFICIENT
RS28707189	2566752	1	7.82E-01
RS8058306	67440289	2	7.12E-01
RS2734743	4937909	3	6.57E-01
RS12598147	81995177	4	-6.08E-01
RS16965197	62855866	5	5.50E-01
RS9930906	2534525	6	5.29E-01
RS9932300	7939088	7	5.22E-01
RS4523927	2471582	8	4.42E-01
RS4122247	49476172	9	4.39E-01
RS17136933	4288958	10	4.36E-01

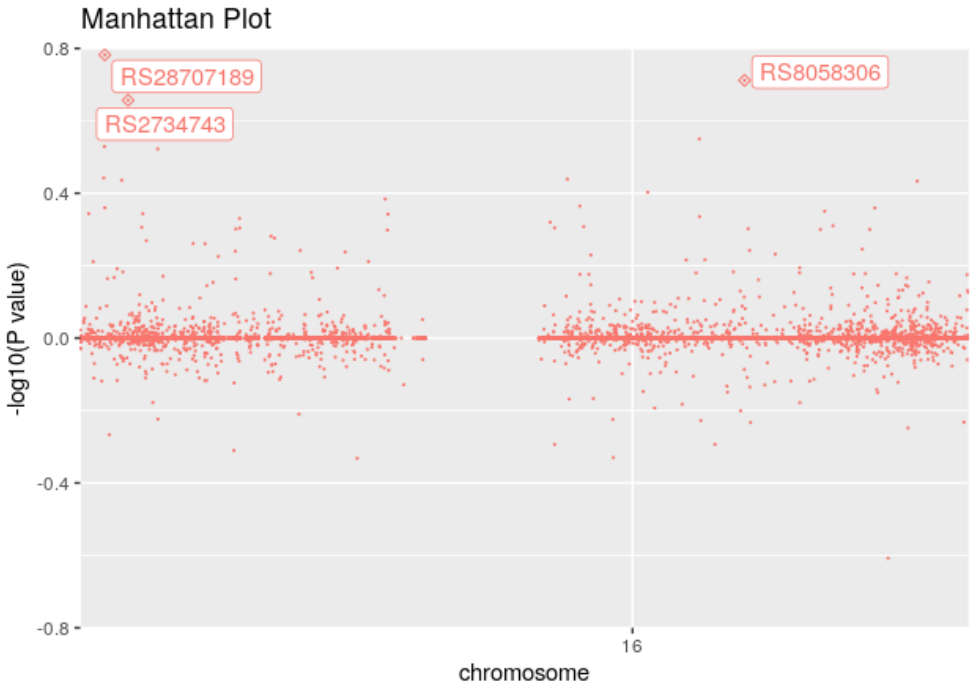

Figure 28: Manhattan Plot of the top three SNPs annotated for lasso method.

3.6.5 Results - Frequentist Method

Results for the top ten SNPs ranked by the frequentist method revolve around one spike of SNPs located close together. It can be seen from Table 41, Figure 30 and Table 42 that blocks of SNPs that are 3rd, 4th, 5th and 6th (***) ranked by the frequentist method, are in near perfect linkage disequilibrium with each other. The block of four SNPs (**), rs7206295, rs4597342, rs8060268 and rs4075052 also share high linkage disequilibrium. Another block shows SNPs rs9888879 and rs35314490 (*) are in perfect linkage disequilibrium. All blocks have high R^2 scores with each other, meaning they are highly correlated.

Table 41: Top ten SNPs ranked for Chromosome 16 and accompanied with their p-value

Frequentist			
SNP	POSITION	RANK	P-VALUE
RS35314490	31283164	1	1.13E-60
RS9888879	31310372	2	2.69E-60
RS4632147	31363381	3	8.39E-54
RS1143678	31343005	4	4.59E-51
RS1143683	31336888	5	2.04E-50
RS11574637	31368874	6	4.90E-49
RS4548893	31364493	7	3.21E-41
RS9937837	31298939	8	4.75E-29
RS11644034	**85972612**	**9**	**1.17E-21**
RS13332545	31377390	10	2.85E-19

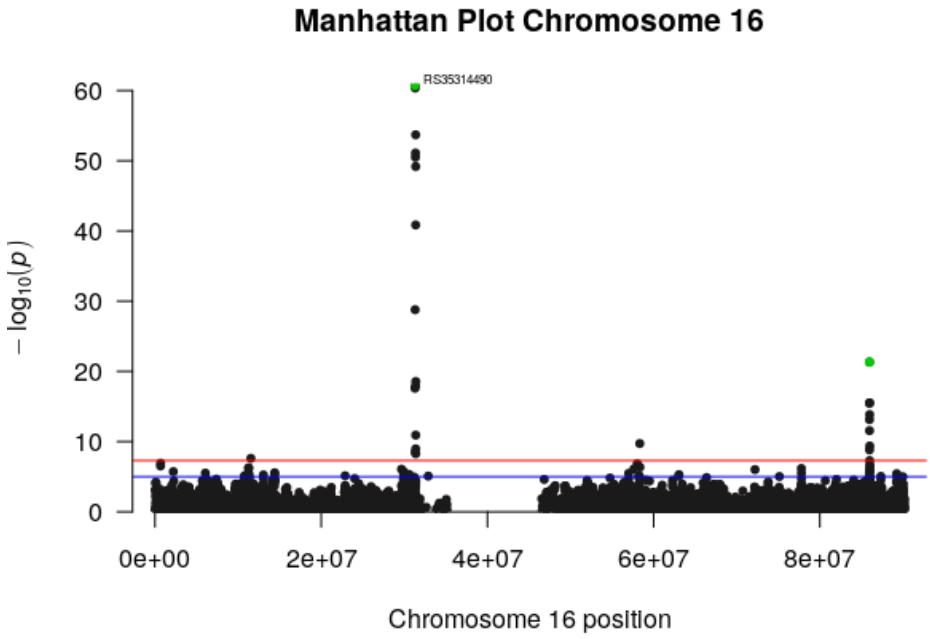

Figure 29: A Manhattan plot with SNPs rs35314490 annotated and rs11644034 highlighted with a green spot

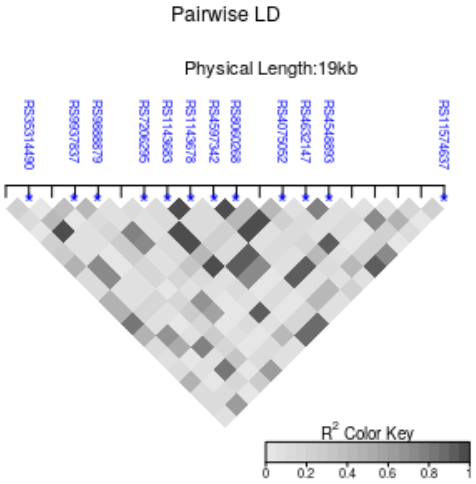

Figure 30: Strong linkage disequilibrium amongst highly ranked SNPs

Table 42: Chromosome 16 - A block of SNPs showing position, result and LD.

Chromosome 16 from Position 31281147-31367318						
SNP	LD	Position	SPIKE	LASSO	FREQ	EBEN
RS2359661		31281147	0	0	2.68E-18	8.90E-02
RS35314490	*	31283164	5.80E-01	3.42E-01	1.13E-60	2.34E-02
RS11861251		31289396	0	0	1.03E-02	0
RS9937837		31298939	0	1.00E-02	4.75E-29	1.24E-02
RS9888879	*	31310372	0	4.00E-02	2.69E-60	7.00E-02
RS11645653		31312855	-7.00E-02	-6.00E-02	4.39E-09	1.36E-01
RS7206295	**	31336519	0	0	2.81E-04	0
RS1143683	***	31336888	0	0	2.04E-50	0
RS1143678	***	31343005	0	0	4.59E-51	0
RS4597342	**	31343769	0	0	1.59E-04	0
RS8060268	**	31345280	0	0	3.42E-04	0
RS3925075		31347748	0	0	8.69E-19	0
RS4075052	**	31348233	0	0	1.58E-04	0
RS4632147	***	31363381	0	-9.00E-02	8.39E-54	4.48E-02
RS4548893		31364493	0	-6.05E-04	3.21E-41	1.94E-01
RS11863903		31364909	0	0	1.39E-04	0
RS11150614		31366016	0	0	6.90E-05	0
RS11574633		31367318	0	0	2.71E-03	0
RS7190997		31368178	0	0	1.83E-09	0
RS11574637	***	31368874	0	0	4.89E-49	4.82E-01

3.6.6 Results - EBEN Method

All SNPs failed to make genome wide significance.

Table 43: Top ten SNPs ranked for Chromosome 16 and accompanied with their p-value

EBEN			
SNP	POSITION	RANK	P-VALUE
RS9926690	79638121	1	4.57E-05
RS305059	85976018	2	1.79E-04
RS11640961	30979818	3	2.26E-04
RS3764261	56993324	4	2.75E-04
RS17829520	84724295	5	3.11E-04
RS12596171	24338109	6	3.28E-04
RS158481	57075253	7	3.46E-04
RS2288012	58327646	8	4.02E-04
RS1992893	10125302	9	5.62E-04
RS1473204	60514331	10	5.87E-04

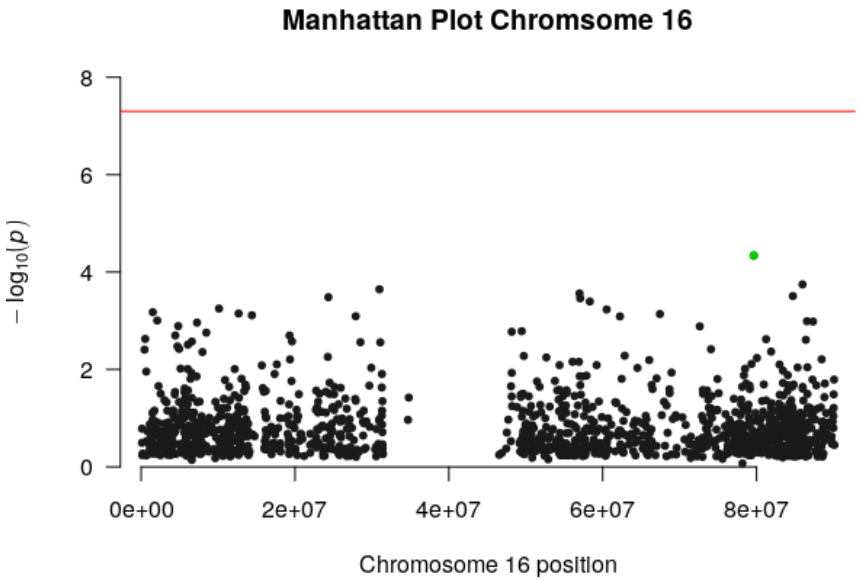

Figure 31: Manhattan Plot of the top SNPs for the lasso method

3.6.7 Results - Summary

For chromosome 16, from 20388 SNPs, the spike method produced 1272 non-zero variables, lasso method produced 1772 and EBEN method produced 1303. The study found the associated SNP rs11574637 but only with frequentist and EBEN methods. This SNP has previously been acknowleged to be a risk locus for myeloid leukemia [116], and IgA nephropathy [117] in European populations. SNP rs35314490 was found by all four methods and was the top ranked SNP by the frequentist method (see Figure 45 and Table 44) although no previous associations have been found with any disease.
The associated SNP rs9888739 found by Harley et al [41] in 2008, and SNP rs1170426 found by Morris et al [43] in 2016 in Asian and European populations, were not included in this study. Associated SNPs rs9652601 and rs34572943 were both imputed in the original study by Bentham et al and were also not part of this study.

Associated SNP in Chinese populations, rs12599402 [118], ranked 136th in the frequentist method but had a zero-coefficient with the variable selection methods.

The associated SNP rs11644034 reported by Bentham et al in the gene *IRF8* was found by all four methods but it was only ranked in the top ten in the frequentist method. Bentham et al found SNP rs2288012 which was chosen by all four methods and featured in the top 20 ranked in two methods (frequentist and EBEN).

SNP rs8058306 features highly amongst three of the methods but has no recorded associations with lupus but a recent study found a link between the SNP and variants affecting bone mineral density in Hispanic people [119]. SNP rs3764261 has been linked with cholesterol problems [120] and is ranked in the top 160 by all four methods but no previous associations have been made with lupus. Langefeld et al study [63] in Hispanic Americans made an association between SNPs rs2550333, rs2731763 (in gene *CCDC113*) and SLE. The variable selection methods produced zero-coefficients but the frequentist method made the SNP rs2550333 as the 158th top ranked SNP. The SNP rs 2731763 which has a D'=1 and $r^2 = 0.848$ with rs2550333 and a distance close to 25Kb was found in this study by all four methods with robust results.

Table 44: Chromosome 16 - Top ten SNPs ranked for each method

Chromosome 16					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS28707189	2566752	1	1	3465	0
RS8059824	67963284	2	0	4748	0
RS16957597	67946356	3	47	1483	0
RS13335252	67808212	4	32	4834	0
RS8058306	67440289	5	2	18255	12
RS9930906	2534525	6	6	12713	136
RS7187296	3036551	7	42	2873	171
RS16965197	62855866	8	5	14639	41
RS1486445	51857372	9	108	4917	0
RS2734743	4937909	10	3	17609	36
RS12598147	81995177	11	4	92	0
RS9932300	7939088	12	7	339	0
RS4523927	2471582	16	8	15724	0
RS4122247	49476172	28	9	4684	22
RS17136933	4288958	22	10	1593	0
RS35314490	31283164	46	20	1	110
RS9888879	31310372	0	446	2	252
RS4632147	31363381	0	188	3	178
RS1143678	31343005	0	0	4	0
RS1143683	31336888	0	0	5	0
RS11574637	31368874	0	0	6	1103
RS4548893	31364493	0	1705	7	418
RS9937837	31298939	0	1152	8	70
RS11644034	**85972612**	**567**	**461**	**9**	**698**
RS13332545	31377390	0	771	10	868
RS9926690	79638121	132	223	359	1
RS305059	85976018	108	200	17	2
RS11640961	30979818	87	420	16902	3
RS3764261	56993324	131	160	44	4
RS17829520	84724295	120	210	167	5
RS12596171	24338109	109	152	5202	6
RS158481	57075253	184	521	12313	7
RS2288012	58327646	111	164	19	8
RS1992893	10125302	123	403	525	9
RS1473204	60514331	116	173	13644	10

Table 45: Chromosome 16 - Top ten SNPs for each method

Chromosome 16				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS28707189	2.64E+00	7.82E-01	5.82E-02	0
RS8059824	1.50E+00	0	9.68E-02	0
RS16957597	1.34E+00	2.42E-01	1.27E-02	0
RS13335252	1.31E+00	3.02E-01	9.98E-02	0
RS8058306	1.23E+00	7.12E-01	8.56E-01	7.28E-04
RS9930906	1.18E+00	5.29E-01	4.91E-01	3.15E-02
RS7187296	-1.01E+00	-2.67E-01	4.21E-02	4.30E-02
RS16965197	9.87E-01	5.50E-01	6.17E-01	5.22E-03
RS1486445	9.84E-01	1.47E-01	1.02E-01	0
RS2734743	9.68E-01	6.57E-01	8.13E-01	3.78E-03
RS12598147	-9.53E-01	-6.08E-01	2.35E-05	0
RS9932300	9.24E-01	5.22E-01	5.70E-04	0
RS4523927	8.28E-01	4.42E-01	6.88E-01	0
RS4122247	6.40E-01	4.39E-01	9.52E-02	1.64E-03
RS17136933	7.12E-01	4.36E-01	1.48E-02	0
RS35314490	5.78E-01	3.42E-01	1.13E-60	2.34E-02
RS9888879	0	4.26E-02	2.69E-60	7.00E-02
RS4632147	0	-8.92E-02	8.39E-54	4.48E-02
RS1143678	0	0	4.59E-51	0
RS1143683	0	0	2.04E-50	0
RS11574637	0	0	4.90E-49	4.82E-01
RS4548893	0	-6.05E-04	3.21E-41	1.94E-01
RS9937837	0	1.46E-02	4.75E-29	1.24E-02
RS11644034	**-1.99E-02**	**-4.08E-02**	**1.17E-21**	**2.45E-01**
RS13332545	0	2.33E-02	2.85E-19	3.45E-01
RS9926690	7.33E-02	7.67E-02	6.45E-04	4.57E-05
RS305059	1.06E-01	8.54E-02	3.20E-12	1.79E-04
RS11640961	4.26E-01	4.54E-02	7.13E-01	2.26E-04
RS3764261	7.41E-02	1.00E-01	3.22E-06	2.75E-04
RS17829520	8.84E-02	8.10E-02	1.26E-04	3.11E-04
RS12596171	1.04E-01	1.07E-01	1.13E-01	3.28E-04
RS158481	-5.36E-02	-3.69E-02	4.68E-01	3.46E-04
RS2288012	1.01E-01	9.86E-02	1.98E-10	4.02E-04
RS1992893	-8.74E-02	-5.97E-02	1.51E-03	5.62E-04
RS1473204	-9.23E-02	-9.37E-02	5.53E-01	5.87E-04

3.7 Chromosome 22

3.7.1 Introduction

Chromosome 22 is the second smallest in the whole genome containing around 49 million base pairs.

3.7.2 Previous GWAS associations in Chromosome 22

Previously, this chromosome has only one associated risk allele to lupus in European population, rs7444 (see Table 46). This marker first came to attention in 2012 in a paper by Wang et al [121], they concluded that in a functional haplotype of *UBE2L3*, rs7444 was associated with the disease lupus. Bentham et al [4] replicated this in Europeans, resulting in a post–replication study meta–analysis p-value of 1.84E-22. Associated SNP rs131654 was found in a study by Han et al [102] in an East Asian population, to have an association to lupus.

Table 46: Timeline of asscociated SNPs with lupus through GWAS 2008-2018.

GWAS 2008-2018					
Year	Chr	Associated SNP	Likely causal gene	Study population	Author
2012	22	rs7444	*UBE2L3*	EC	WAN

KEY: EUR=European, EC=European and Chinese, EA=European American, WAN=Wang et al [121].

3.7.3 Results - Spike and Slab Method

Table 47: Top ten SNPs ranked for Chromosome 22 and accompanied with their coefficient

Spike and Slab			
SNP	POSITION	RANK	COEFFICIENT
RS7285053	48872193	1	2.48E+00
RS7292957	37057019	2	1.96E+00
RS4821348	35446992	3	1.53E+00
RS6000370	37056207	4	1.48E+00
RS138070	44229508	5	-1.37E+00
RS1800706	19928022	6	1.27E+00
RS9605031	19921378	7	1.27E+00
RS138065	44226987	8	1.19E+00
RS12484382	35443638	9	-1.08E+00
RS204970	34871591	10	-1.07E+00

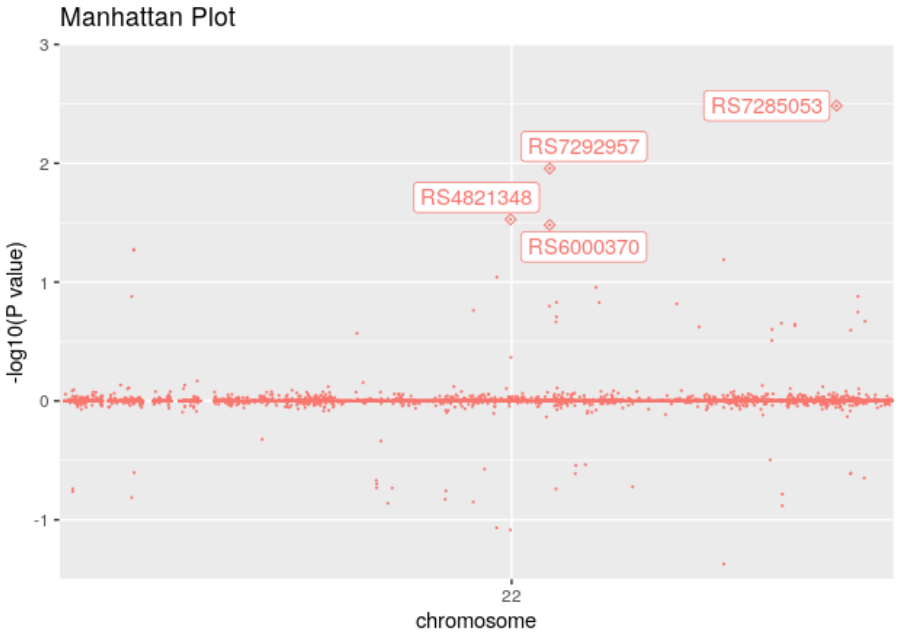

Figure 32: A Manhattan plot with the top four SNPs annotated for the spike method

3.7.4 Results - Lasso Method

The lasso method's data was again similar to that of the spike method in SNPs chosen that were highly ranked. The lasso method produced 3242 (33.9%) of non-zero coefficients, the highest seen across the whole genome. Compared to the other methods, this was extremely high with no apparent reason for this figure (see Table 9).

3.7.5 Results - Frequentist Method

SNPs that were ranked 2-11 by the frequentist method are positionally tightly packed within a block of 10 SNPs. Figure 35 represents this block indicating the R^2 calculations. It is worth noting that the other methods produced mostly zero coefficients for this specific group.

3.7.6 Results - EBEN Method

The associated SNP rs7444 returned a rank of 6th using the EBEN method. As per the other methods the top SNP was rs7285053.

Table 48: Top ten SNPs ranked for Chromosome 22 and accompanied with their coefficient using the spike method.

Lasso			
SNP	POSITION	RANK	COEFFICIENT
RS7285053	48872193	1	3.26E+00
RS7292957	37057019	2	1.77E+00
RS6000370	37056207	3	1.46E+00
RS138070	44229508	4	-1.29E+00
RS138065	44226987	5	9.48E-01
RS9618690	19777791	6	8.92E-01
RS7287114	37040092	7	8.81E-01
RS6002526	42289565	8	8.39E-01
RS17002469	41856041	9	-7.87E-01
RS6002958	43218614	10	7.50E-01

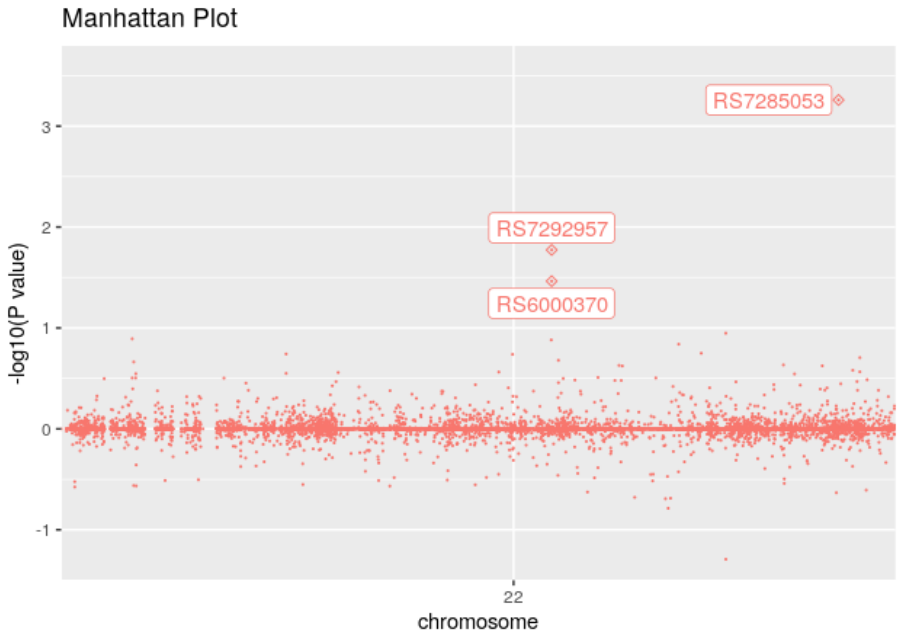

Figure 33: Manhattan Plot of the top SNPs for the lasso method

3.7.7 Results - Summary

For chromosome 22, from 9563 SNPs, the spike method produced 1170 non-zero variables, lasso produced 3242 and EBEN produced 1111. From Table 51, the SNP rs7285053 was the only SNP in this study that was ranked first for all methods. This SNP has no previous acknowledgement of any association with

Table 49: Top ten SNPs ranked for Chromosome 22 and accompanied with their p-value using data from the frequentist method.

Frequentist			
SNP	POSITION	RANK	P-VALUE
RS7285053	48872193	1	7.42E-27
RS7444	**21976934**	**2**	**3.84E-13**
RS11089637	21979096	3	4.13E-13
RS5754217	21939675	4	5.76E-13
RS878825	21982249	5	9.92E-13
RS181360	21928916	6	8.07E-12
RS1034329	21943938	7	1.01E-11
RS140498	21927064	8	1.12E-11
RS4821116	21973319	9	1.78E-11
RS5998619	21945851	10	1.97E-11

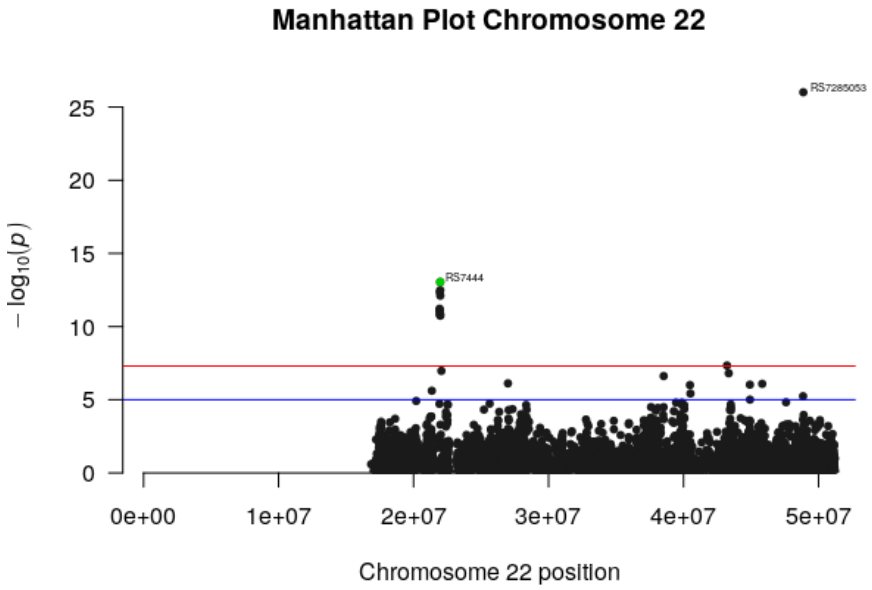

Figure 34: A Manhattan plot with SNPs rs7444 and rs7285053 annotated from the frequentist data

any disease. SNPs that ranked from 5-10 by the frequentist method, all had received a zero coefficient from the other three methods. All the SNPs ranked in the top ten by the EBEN method were also ranked highly in the other variable selection methods but not by the frequentist method (barring rs7285053).

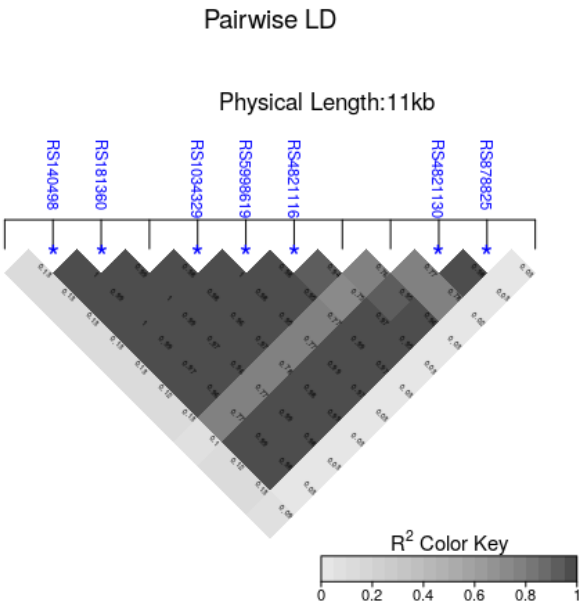

Figure 35: A LD heatmap of a dense block of ten strongly correlated SNPs that ranked in the top 11 by the frequentist method.

Table 50: Top ten SNPs ranked for Chromosome 22 and accompanied with their coefficient using data from the EBEN method.

EBEN			
SNP	POSITION	RANK	P-VALUE
RS7285053	48872193	1	3.57E-12
RS9605179	17412216	2	3.69E-12
RS9606542	17410658	3	8.17E-11
RS4821348	35446992	4	2.25E-10
RS8135828	29929239	5	9.27E-09
RS7444	**21976934**	**6**	**2.52E-08**
RS5756407	37316259	7	2.99E-08
RS17344701	41132402	8	1.52E-07
RS9614670	45838817	9	6.90E-07
RS11704508	46638540	10	1.12E-06

This group of SNPs have no recorded association with lupus apart from rs7444 although rs878825 has links to red cell distribution [122] and rs4821116 links with HDL cholesterol levels [123]. SNP rs6000370 was found by all four methods ranking in the top 14 by methods spike, lasso and EBEN, although there has been no association recorded with any disease. Associated SNP rs131654 was

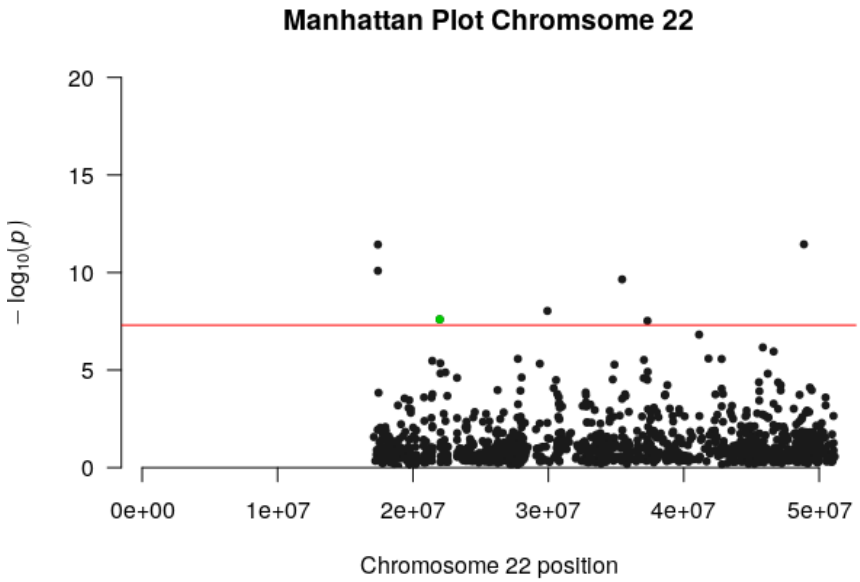

Figure 36: A Manhattan plot with the associated SNP rs7444 highlighted for the EBEN method

found in a study by Han et al [102] in 2009 in an East Asian population, to have an association to lupus. This SNP was weakly represented by methods spike and lasso and ranked 27th by the frequentist method, although the EBEN produced a zero coefficient. Associated SNP rs7444 was found by all four methods with strong results.

Table 51: Chromosome 22 - Top ten SNPs ranked for each method

Chromosome 22					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS7285053	48872193	1	1	1	1
RS7292957	37057019	2	2	127	476
RS4821348	35446992	3	11	9244	4
RS6000370	37056207	4	3	76	14
RS138070	44229508	5	4	58	0
RS1800706	19928022	6	37	219	0
RS9605031	19921378	7	42	1147	928
RS138065	44226987	8	5	9054	318
RS12484382	35443638	9	697	9325	59
RS204970	34871591	10	69	7245	269
RS9618690	19777791	0	6	3873	0
RS7287114	37040092	23	7	2156	25
RS6002526	42289565	21	8	1592	185
RS17002469	41856041	0	9	4805	0
RS6002958	43218614	43	10	12	1095
RS7444	**21976934**	**76**	**485**	**2**	**6**
RS11089637	21979096	0	998	3	0
RS5754217	21939675	83	2167	4	0
RS878825	21982249	0	0	5	0
RS181360	21928916	0	0	6	0
RS1034329	21943938	0	0	7	0
RS140498	21927064	0	0	8	0
RS4821116	21973319	0	0	9	0
RS5998619	21945851	0	0	10	0
RS9605179	17412216	25	27	7891	2
RS9606542	17410658	29	40	7104	3
RS8135828	29929239	32	43	1678	5
RS5756407	37316259	30	65	2624	7
RS17344701	41132402	121	68	1105	8
RS9614670	45838817	69	87	7621	9
RS11704508	46638540	24	53	2260	10

Table 52: Chromosome 22 - Top ten SNPs for each method

Chromosome 22				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS7285053	2.48E+00	3.26E+00	7.42E-27	3.57E-12
RS7292957	1.96E+00	1.77E+00	5.73E-04	1.00E-01
RS4821348	1.53E+00	7.38E-01	9.53E-01	2.25E-10
RS6000370	1.48E+00	1.46E+00	2.49E-04	3.03E-06
RS138070	-1.37E+00	-1.29E+00	1.36E-04	0
RS1800706	1.27E+00	5.46E-01	1.72E-03	0
RS9605031	1.27E+00	5.05E-01	3.59E-02	4.02E-01
RS138065	1.19E+00	9.48E-01	9.28E-01	4.19E-02
RS12484382	-1.08E+00	-1.08E-01	9.64E-01	2.96E-04
RS204970	-1.07E+00	-4.50E-01	6.72E-01	2.90E-02
RS9618690	0	8.92E-01	2.63E-01	0
RS7287114	7.98E-01	8.81E-01	1.03E-01	2.59E-05
RS6002526	8.18E-01	8.39E-01	6.23E-02	1.24E-02
RS17002469	0	-7.87E-01	3.61E-01	0
RS6002958	6.22E-01	7.50E-01	4.31E-08	5.84E-01
RS7444	**1.03E-01**	**1.45E-01**	**3.84E-13**	**2.52E-08**
RS11089637	0	7.99E-02	4.13E-13	0
RS5754217	-9.37E-02	-3.07E-02	5.76E-13	0
RS878825	0	0	9.92E-13	0
RS181360	0	0	8.07E-12	0
RS1034329	0	0	1.01E-11	0
RS140498	0	0	1.12E-11	0
RS4821116	0	0	1.78E-11	0
RS5998619	0	0	1.97E-11	0
RS9605179	-7.63E-01	-5.76E-01	7.56E-01	3.69E-12
RS9606542	-7.40E-01	-5.23E-01	6.52E-01	8.17E-11
RS8135828	-7.31E-01	-5.11E-01	6.88E-02	9.27E-09
RS5756407	-7.40E-01	-4.61E-01	1.43E-01	2.99E-08
RS17344701	-6.96E-02	-4.51E-01	3.40E-02	1.52E-07
RS9614670	-1.18E-01	-3.93E-01	7.21E-01	6.90E-07
RS11704508	-7.84E-01	-4.96E-01	1.12E-01	1.12E-06

From Table 52, it is plain to see, the EBEN method has combined all the linkage disequilibrium and produced just one SNP to represent this correlation (rs7444). Whilst the Spike method has been a little less certain and spread its correlated SNPs into two but with a slightly weaker result (rs7444 and rs5754217). Furthermore, the lasso method has been even looser and allowed three of the SNPs to be representative of the block with SNPs rs7444, rs5754217 and rs11089637 compensated with a weaker beta coefficient score. Figure 37 below presents a larger area of SNPs including the correlated alleles.

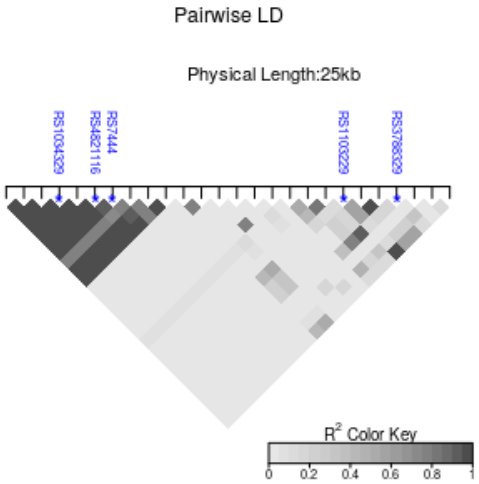

Figure 37: A larger area of SNPs containing the dense block of correlated SNPs that was displayed above.

4 Conclusion and Discussion

In this section it is discussed how the methods performed and what more general conclusions can be made. The frequentist and EBEN methods are presented individually but the spike and lasso have been presented together as they produced very similar results. The comparisons of the methods applied are noted, with what we can learn from the results and any future work moving forwards.

4.1 Introduction

In the analysis of large scale genetic datasets, advanced statistical techniques of variable selection are designed to reduce the number of predictors included in the model of disease susceptibility and hence we can reduce the number of false positives. It is challenging to do this in a realistic way using simulated data and so this work presents an insight into how these methods work with real data. In this thesis, four methods were applied to the post-cleaned dataset from Bentham et al [4]. The data involved 4036 SLE cases and 6959 healthy controls totalling 644,674 SNPs. SNPs that are inherited together in the same biological pathway tend to be correlated and thus variable selection methods were chosen to see how they deal with this. Disease association with potential risk alleles was assessed by means of different advanced statistical methods of regularization using Bayesian and frequentist techniques. Correlated genetic markers were grouped, measured and studied, with the aspiration of finding positive association with SLE. The results were put into tabular form with rankings and their p-values or beta coefficients. Comparisons were made with 38 associated SNPs that were found on the dataset from Bentham et al across the whole genome barring the sex chromosomes.

4.2 Spike and Lasso

The lasso [85] and spike [86], methods adopt lasso as their regularization techniques. Regularization is required for problems with fitting and interpretability to produce an optimal model. The lasso uses a regressor selecting process that either reduces the coefficent to zero or classes it as important and keeps it in the model. The spike method that incorporated a double exponential prior on coefficients is known to act like lasso penalized regression. By introducing a mixture model into the prior this technique goes beyond a standard lasso. The spike method used the same software for the lasso method including choosing the optimal penalty value (glmnet) [90] which was further processed using the BhGLM package [124] using a logistic regression with the spike and slab double exponential prior. Although the results obtained could possibly be viewed as similar in which non-zero coefficients were included, the spike method has been more severe in the reduction of parameters. This was due to the predictors being divided into groups with the assumption that they follow the spike and slab priors, with the main spike around zero and the diffuse distribution for the

slab. There was a steady increase in the percentage of non-zero coefficients that the spike method chose as the chromosomes got smaller in size. This relationship appears to carry through to all methods. The technique of the lasso that did not involve the extra BMlasso software package of the spike method [86], produced a less complex model, consequently, culminating in a less sparse set of SNPs. The spike results were a reduced replication of the lasso's. The six chromosomes featured in the main results section produced an insight into the comparatively similar product of both lasso style methods. Covering the 60 SNPs that featured in the top ten ranked per chromosome by the spike method, 32 of the same SNPs were featured in the top ten of the lasso. In general, for each SNP chosen by the spike, it was very likely the lasso had chosen it too. The spike and slab appears to have over-shrunk the variables too much in favour of sparsity over accuracy or potentially has a poorly chosen prior that leads to predictions that are strongly incorrect.

4.3 EBEN

The EBEN method [84] was computationally intensive. The elastic net, through the l_1, lasso penalty part, uses variable selection and the l_2, the ridge penalty part, produces better prediction accuracy. It has used groups of correlated variables and produced sparse efficient predictors. The elastic net has been made to control the within group correlations. The EBEN's most extreme results feature in chromosome 3, where there is a stark difference in non-zero SNPs chosen compared to the other variable selection methods. Of the top 30 most extreme p-values produced by the EBEN method, 28 appear in chromosome 3. The SNPs rs10148260 in chromosome 14 (2.40E-14) and rs9605179 (3.69E-12) in chromosome 22 are the other two. 320 SNPs were recorded at the Bonferroni correction of 5×10^{-8} or less and they are not positioned close to each other. The reason for this extremity is unexplained but is possibilty related to the complexity of the model that produced extended computational processing time compared to the other methods. In comparison to chromosome 6 that had around the same amount of SNPs tested (42483 compared with 42993) the resulting non-zero variables were a contrast to that of chromosome 3. For chromosome 3 totalling 7848 SNPs chosen (compared to chromosome 6 of 488). A theory might be that this is due to the high linkage disequilibrium blocks that are contained in the MHC. This could bring into question EBEN's technique of how to deal with low linkage disequilibrium SNPs.

4.4 Frequentist Method

Produced using SNPTest software [12], the frequentist method ran association tests on the data from Bentham et al study [4]. This technique with no variable selection chose the most SNPs previously associated with lupus from the 2015 paper. 21 hits ranking in the top 25 from the 38 risk alleles and 26 hits from

38 were ranked in the top 100. Although finding the most previously reported hits of the four methods, the results included many false positives. For many chromosomes, the frequentist method had chosen high ranking SNPs that are positioned in local groups close to each other where high linkage disequilibrium exists. This is repeated time and again for each chromosome. It is also noticeable when the variable selection methods have a high ranking, the frequentist method does not.

4.5 Overview

In conclusion, noting the small number of non-zero coefficients chosen in chromosome 6 and the aforementioned correlated block of 10 SNPs in chromosome 22, there is some belief that the EBEN method is stricter with correlated blocks of SNPs than the other two variable selection methods. Also, from the results of high ranking SNPs, the EBEN predicts with better accuracy. The spike method has been more severe in variable selection and with less predictive accuracy with known associated SNPs. The frequentist method ranks blocks of correlated SNPs that incorporate the associated risk allele highly, resulting in many false positives. In general, the results were found to be inconsistent with known associations although a disadvantage of using real data is that the true associations are unknown and so it could not be clarified whether the SNPs with significant results were in fact true or false positives. This would require further analysis from bioscientists. The results on chromosome 22 around rs7444 are encouraging, as this is a well established association where the SNP is known to be causal. The frequentist method selects multiple SNPs that are in close LD with rs7444, but the EBEN method only selects this SNP, while the spike and slab, and lasso methods select one or two other SNPs respectively.

Although many SNPs do not pass GWAS significant thresholds, it is likely that some of these SNPs are associated with the disease and further research will be required to produce associations in the future. Overall, the methods show promise statistically, as they generally identify known associations with lupus. However, the methods all clearly identify many false positives, and so could not be used currently in a practical genome-wide association study. Further work will be needed to refine these methods, so that they are producing results that can be replicated robustly. Given this outcome, it is not clear which of the potential novel associations are genuinely linked with lupus and which are false positives. It is therefore not possible to draw any conclusions about the biological implications of the results or about which method might have superior performance in terms of sensitivity or specificity. To the author's knowledge, this is the first time that these methods have been applied to a practical genetics dataset, and so this study cannot be compared with previous findings. This work is the first step in statistical analysis, which will need to be followed by further statistical work, then validation using multiple large genetic datasets, before finally immunological work can be carried out to show the biological effects of the genuine associations with the disease. The very large number of associations

found across the four methods mean that many must be false positives. It would therefore not be sensible to start to carry out bioinformatics studies to assess which genes each of these many hundereds of SNPs are in. This means that it is not possible to assess whether novel genes have been found. Summarising the results into a single plot would also not be appropriate since we know that many of the associations are false positives.

4.6 Future work

An area for future work involves the rising percentage of non-zero coefficients chosen per chromosome as the chromosomes get smaller in size. In addition, an area to further understand is the rationale of EBEN's selection procedure for blocks with high and low linkage disequilibrium. Overall, it is clear to see that other techniques of variable selection should be researched and utilized in place of standard association testing for genetic data. Variations of the group lasso, with sparse and overlap group lasso in particular, could potentially be more consistent with blocks of correlated data. With this in mind, the goal is to find a more consistent model producing reduced number of false positives, that conclude in novel associated hits and will help in the search for the missing heritability in the fight against disease. This work is part of the long process of unravelling the human genome in order to treat complex diseases such as lupus.

A Appendix: Software

Appendix A presents the two main pieces of genetics software used in this thesis with the individual file names that they produced.

A.1 Files for Genetic Software

The main pieces of software used to produce this thesis was SNPTest [12] and PLINK [11].

PLINK v.1.07 (Shaun Purcell http://pngu.mgh.harvard.edu/purcell/plink/)
.bim (txt) file contains genetic markers
Chromosome Number/SNP/Genetic Distance/Base pair postion/Allele 1/Allele 2
.fam (txt) file contains information on the individuals
Family ID/Individual ID/Paternal ID/Maternal ID/Sex/Phenotype
.bed file (Binary fileset) contains binary information
.map file (txt) file contains
Chromosome Number/SNP/Genetic Distance/Base pair postion
.ped file
Family ID/Individual ID/Paternal ID/Maternal ID/Sex/Phenotype/Allele call (1st var)/Allele call (1st var)/Allele call (2nd var)/Allele call (2nd var)
The bim, fam and bed files accompany each other as does the map accompanies the ped files.

SNPTest v.2.4.1. (Jonathon Marchini and Gavin Band)
.gen (Genesis ROM) file contains genotype data for the cohort
Chromosome/SNP/Base pair position/Allele 1/Allele 2
.sample (txt) file contains the IDs and phenotype information
Family ID 1/Within family ID 2/Missing call frequency/Sex/Phenotype
A .gen file should be accompanied by a .sample file

B Appendix: Results

Appendix B relays the rest of the information from the results that do not appear in the results section. These chromosomes have fewer known associations with SLE.

B.1 Chromosome 3

For chromosome 3, from 42483 SNPs, spike method produced 1443 non-zero variables, lasso method produced 1645 and EBEN method produced 7848.

SNP rs9311676 [4] produced 3 zero-coefficients and a weak p-value for frequentist.

SNP rs564799 (Bentham et al) ranked 9th in the frequentist method and had a weak EBEN p-value. Spike and lasso methods produced zero beta values. The closest in distance to SNP rs564799 is rs574808 and this has the spike, lasso and frequentist methods ranked in their top 200.

The frequentist method ranked 33rd with SNP rs10936599 that is situated in gene *MYNN*. The other three methods produced zero-coefficients. This SNP is an associated SNP from Molineros study of Asian and Europeans also Wens study in Chinese population [125].

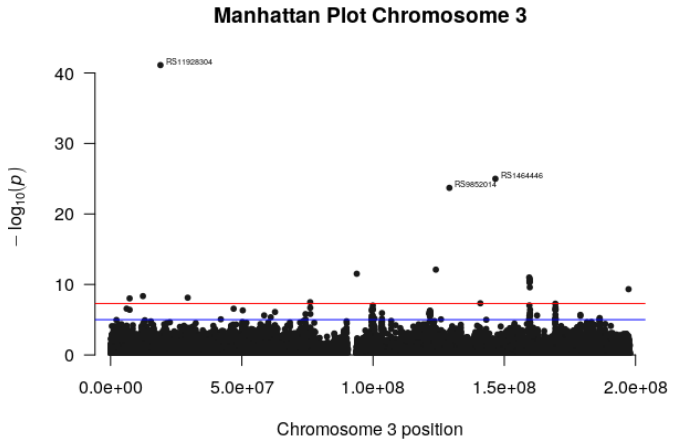

Figure 38: A Manhattan plot with the lowest p-value SNPs rs9852014, rs1464446 and rs11928304 highlighted

Table 53: Chromosome 3 - Top ten SNPs ranked for each method

Chromosome 3 - Top ten SNPs ranked for each method					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS9852014	129084581	1	1	3	0
RS6810203	129083281	2	3	38480	0
RS35755212	197346971	3	4	12	480
RS6414259	93791390	4	5	5	387
RS11921784	118468553	5	2	393	0
RS1464446	146601295	6	7	2	0
RS35992	74267575	7	44	86	0
RS4688732	50360802	8	12	20743	272
RS11928304	18998569	9	6	1	0
RS4318522	146594746	10	8	28130	0
RS6779649	179076879	11	9	10543	11
RS9811883	43121691	14	10	735	384
RS1444766	123925271	135	107	4	0
RS6441275	159501673	193	183	6	3390
RS574808	159732983	143	167	7	0
RS1675497	159582382	0	0	8	0
RS564799	**159728987**	**0**	**0**	**9**	**994**
RS26298	159611291	0	0	10	0
RS9827067	24808506	0	0	32797	1
RS7642268	60084924	0	0	32450	2
RS4234583	1051760	105	248	34185	3
RS6796963	196662275	0	0	10285	4
RS6797560	196620294	0	0	25086	5
RS796313	12449528	0	0	13893	6
RS6787169	149211986	65	26	20201	7
RS2177044	146601081	66	58	39691	8
RS9860651	153416484	145	118	1266	9
RS7618684	57545993	0	1509	17990	10

Table 54: Chromosome 3 - Top ten SNPs for each method

Chromosome 3				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS9852014	-2.96E+00	-1.61E+00	2.75E-25	0
RS6810203	-2.89E+00	-1.44E+00	8.75E-01	0
RS35755212	-2.82E+00	-1.39E+00	4.39E-10	5.26E-06
RS6414259	-1.95E+00	-1.16E+00	6.21E-12	1.44E-06
RS11921784	-1.95E+00	-1.45E+00	3.80E-04	0
RS1464446	-1.88E+00	-9.82E-01	2.22E-25	0
RS35992	1.26E+00	2.16E-01	8.76E-06	0
RS4688732	1.23E+00	4.22E-01	3.61E-01	1.65E-07
RS11928304	-1.02E+00	-1.00E+00	4.55E-42	0
RS4318522	-1.01E+00	-5.34E-01	5.59E-01	0
RS6779649	1.01E+00	4.94E-01	1.24E-01	1.98E-14
RS9811883	-9.08E-01	-4.79E-01	1.17E-03	1.41E-06
RS1444766	-1.25E-01	-1.20E-01	5.16E-12	0
RS6441275	5.68E-02	7.02E-02	1.43E-11	4.69E-02
RS574808	-9.00E-02	-7.38E-02	1.91E-11	0
RS1675497	0	0	3.53E-11	0
RS564799	**0**	**0**	**4.28E-11**	**2.23E-04**
RS26298	0	0	8.02E-11	0
RS9827067	0	0	6.99E-01	2.22E-16
RS7642268	0	0	6.89E-01	4.44E-16
RS4234583	3.74E-01	5.45E-02	7.42E-01	6.66E-16
RS6796963	0	0	1.19E-01	1.33E-15
RS6797560	0	0	4.75E-01	1.78E-15
RS796313	0	0	1.95E-01	4.00E-15
RS6787169	5.37E-01	2.96E-01	3.47E-01	7.33E-15
RS2177044	-5.30E-01	-1.70E-01	9.11E-01	9.77E-15
RS9860651	-8.37E-02	-1.05E-01	3.27E-03	1.51E-14
RS7618684	0	-1.65E-03	2.90E-01	1.71E-14

The EBEN method processed chromosome 3 and returned 226 p-values that were statistically significant. See Table 55 for the top 35 ranked SNPs.

Table 55: Chromosome 3 - Top thirty five SNPs ranked by EBEN

Chromosome 3 - Top thirty five SNPs ranked by EBEN		
SNP	Position	p-value
RS9827067	24808506	2.22E-16
RS7642268	60084924	4.44E-16
RS4234583	1051760	6.66E-16
RS6796963	196662275	1.33E-15
RS6797560	196620294	1.78E-15
RS796313	12449528	4.00E-15
RS6787169	149211986	7.33E-15
RS2177044	146601081	9.77E-15
RS9860651	153416484	1.51E-14
RS7618684	57545993	1.71E-14
RS6779649	179076879	1.98E-14
RS2715671	127128648	2.18E-14
RS2720309	123237517	2.78E-14
RS9818318	58782604	4.04E-14
RS6776465	143296624	8.22E-14
RS4284958	184359263	1.39E-13
RS12493138	184359176	1.91E-13
RS9836757	179642221	2.18E-13
RS7646362	64642313	3.12E-13
RS3774490	53663648	3.35E-13
RS1026572	160887654	4.23E-13
RS4683749	142867543	8.85E-13
RS4928154	98899033	1.42E-12
RS6764014	33980449	1.50E-12
RS335825	196672840	1.64E-12
RS1242075	143366830	1.76E-12
RS12486909	176556131	2.40E-12
RS1687282	14810854	2.88E-12
RS2320958	2780449	3.00E-12
RS11926725	3434942	4.38E-12
RS9855458	27639142	4.45E-12
RS11705763	112856961	5.14E-12
RS12696007	153418189	5.90E-12
RS13098112	29263650	6.68E-12
RS4679644	59351946	8.53E-12

B.2 Chromosome 4

Bentham et al [4] found SNP rs10028805 and was ranked 12th in the frequentist method with a p-value of 1.18e-08. The other methods calculated a zero-coefficient. This SNP had previously been reported to be a risk locus for chronic lymphocytic leukemia [126].

Also in the gene *BANK1*, the SNP rs17266594 that was ranked 8th by the frequentist method, has been associated with SLE [63].

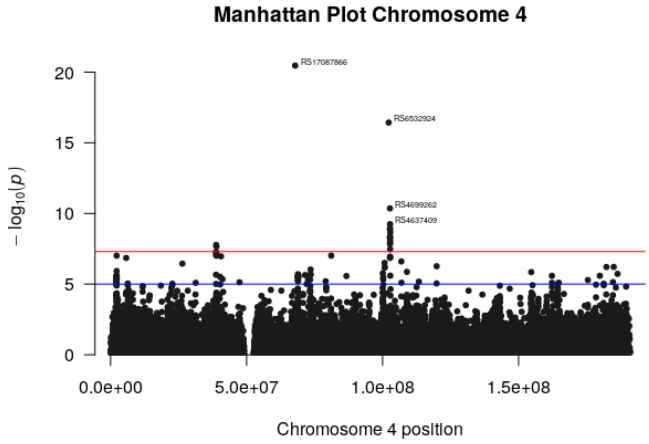

Figure 39: A Manhattan plot with SNPs rs17087866, rs6532924, rs4699262 and rs4637409 highlighted

Table 56: Chromosome 4 - Top ten SNPs ranked for each method

Chromosome 4					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS11466649	38776725	1	107	2516	0
RS9996228	38777517	2	0	29395	0
RS17087866	67876041	3	1	1	1
RS1497941	167058019	4	13	34243	18
RS904246	167044948	5	84	33257	0
RS788927	73364994	6	50	75	0
RS3775152	72630642	7	3	35835	0
RS6828254	110737549	8	6	7371	0
RS6532924	102239170	9	2	2	0
RS11133414	53069867	10	7	4045	0
RS17086015	56555805	12	4	227	0
RS4974786	4315388	22	5	9036	0
RS4585329	115560225	30	8	343	0
RS11569047	110909169	28	9	2172	235
RS17191192	74440419	21	10	4165	0
RS4699262	102726005	0	0	3	72
RS4637409	102753408	185	240	4	463
RS10031210	102714886	295	473	5	127
RS11944613	102718795	0	0	6	0
RS3733197	102839287	0	0	7	0
RS13146194	102723640	0	0	8	0
RS17266594	102750922	0	0	9	0
RS4699258	102710688	0	0	10	0
RS1351357	5746620	213	274	27	2
RS1459455	34595042	171	154	472	3
RS10005935	78682581	179	363	2028	4
RS1458043	81113195	217	432	21	5
RS1263338	2939018	182	148	1250	6
RS10024198	95060602	210	262	576	7
RS1491370	20908415	286	421	223	8
RS17473710	164571285	168	144	252	9
RS12649485	113350665	169	199	81	10

Table 57: Chromosome 4 - Top ten SNPs for each method

Chromosome 4				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS11466649	-1.95E+00	-1.28E-01	1.48E-02	0
RS9996228	-1.61E+00	0	7.47E-01	0
RS17087866	-1.49E+00	-1.21E+00	2.80E-21	1.79E-11
RS1497941	1.30E+00	3.96E-01	9.20E-01	2.69E-02
RS904246	-1.28E+00	-1.47E-01	8.83E-01	0
RS788927	-1.26E+00	-1.84E-01	6.38E-06	0
RS3775152	1.21E+00	6.12E-01	9.78E-01	0
RS6828254	1.14E+00	5.39E-01	8.94E-02	0
RS6532924	1.13E+00	7.26E-01	3.10E-17	0
RS11133414	1.13E+00	4.89E-01	3.31E-02	0
RS17086015	-1.10E+00	-6.11E-01	1.34E-04	0
RS4974786	-8.74E-01	-5.65E-01	1.24E-01	0
RS4585329	7.43E-01	4.43E-01	3.12E-04	0
RS11569047	7.91E-01	4.43E-01	1.15E-02	2.21E-01
RS17191192	8.90E-01	4.20E-01	3.48E-02	0
RS4699262	0	0	4.74E-11	9.14E-02
RS4637409	6.89E-02	6.32E-02	6.22E-10	3.85E-01
RS10031210	-3.87E-02	-3.80E-02	1.30E-09	1.41E-01
RS11944613	0	0	1.50E-09	0
RS3733197	0	0	2.45E-09	0
RS13146194	0	0	4.89E-09	0
RS17266594	0	0	5.35E-09	0
RS4699258	0	0	6.60E-09	0
RS1351357	5.54E-02	5.71E-02	1.67E-07	1.66E-03
RS1459455	7.85E-02	9.20E-02	6.81E-04	4.42E-03
RS10005935	-7.01E-02	-4.47E-02	1.04E-02	5.91E-03
RS1458043	5.35E-02	3.81E-02	1.09E-07	9.04E-03
RS1263338	7.02E-02	9.49E-02	4.55E-03	1.14E-02
RS10024198	5.62E-02	5.96E-02	1.03E-03	1.33E-02
RS1491370	3.98E-02	3.92E-02	1.27E-04	1.36E-02
RS17473710	-9.89E-02	-9.63E-02	1.61E-04	1.37E-02
RS12649485	9.57E-02	7.25E-02	7.85E-06	1.48E-02

Table 58: An area of SNPs in gene *BANK1* around three associated risk alleles, two (underlined) found by Langefeld et al in 2017 and one (bold) found by Bentham et al. The lower block of six SNPs are in high linkage disequilibrium.

Chromosome 4					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS10031210	102714886	-3.87E-02	-3.80E-02	1.30E-09	1.41E-01
RS11944613	102718795	0	0	1.50E-09	0
RS13146194	102723640	0	0	4.89E-09	0
RS4699262	102726005	0	0	4.74E-11	9.14E-02
RS6833249	102726073	0	0	2.18E-01	0
RS13136297	102736456	0	0	1.30E-08	2.60E-01
RS10028805	**102737250**	**0**	**0**	**1.18E-08**	**0**
RS4276281	102746780	0	0	1.36E-07	0
RS17266594	102750922	0	0	5.35E-09	0
RS10516486	102751276	0	0	1.12E-07	4.68E-01
RS4637409	102753408	6.89E-02	6.32E-02	6.22E-10	3.85E-01

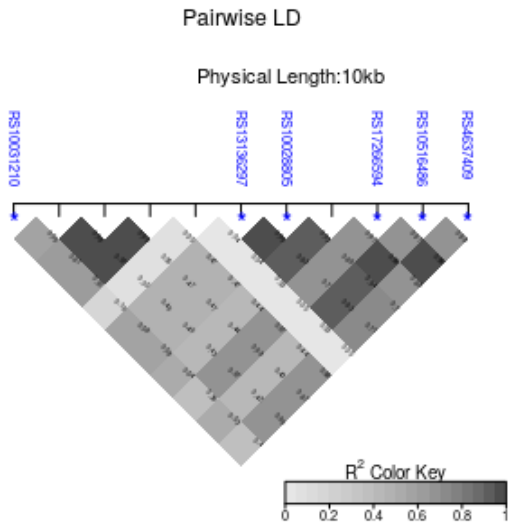

Figure 40: A block of SNPs around three associated risk alleles rs10028805, rs17266594 and rs4637409 in high linkage disequilibrium.

B.3 Chromosome 7

For chromosome 7, from 33932 SNPs, the spike method produced 1366 non-zero variables, lasso method produced 2246 and the EBEN method produced 1154.

Hom et al [7] had observed an association between rs10488631 and lupus. Bentham et al [4] produced significant results too. In this study, all methods produced a rank of 122nd or lower with the frequentist method calculating it as the lowest ranked SNP.

In the gene *IKZF1* Han et al [102] made an association with SNP rs4917014 in Eastern Asian population followed by studies from Bentham et al in Europeans and a mixed population study from Morris et al [47]. All 4 methods chose the SNP ranking between 33rd and 311th.

In a study by Gateva et al [60], SNP rs849142 was found by all 4 methods but weakly.

Associated SNP rs729302 [127] was ranked 13th by frequentist method with a p-value of 4.59E-18 and EBEN producing a p-value of 4.72E-02 and was 17th rank. Another notable hit was rs10954213 that has been associated with Japanese and Korean population [128] ranked 10th by the frequentist method.

In a study from Langefeld et al [63] over a wide population the SNP rs2092540 was found to have an association. The risk locus in the gene *SKAP2* was found by all four methods although weakly.

SNP rs4728142 in the gene *IRF5* was originally observed in a study by Han et al in 2009 in East Asian population followed up by Armstrong et al [62] in 2014 in Europeans. Strong results were observed from all four methods for this SNP.

SNPs associated rs150518861 and rs73135369 by Julia et al [64] and Morris et al respectively, were not found in the study.

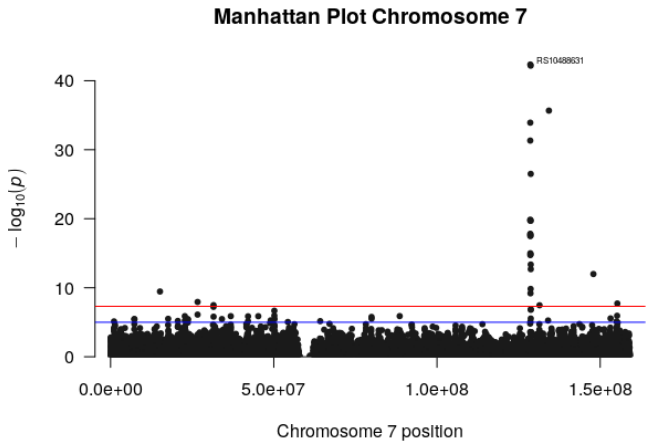

Figure 41: A Manhattan plot with the top ranked SNP 10488631 highlighted

Table 59: Chromosome 7 - Top ten SNPs ranked for each method

Chromosome 7					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS10264693	134289147	1	1	3	1
RS12540166	88573021	2	7	25864	711
RS17880470	94993261	3	29	6906	0
RS17161201	139576284	4	45	10141	0
RS12535041	88573471	5	10	36	0
RS10225000	141728209	6	24	740	0
RS17161202	139578862	7	20	22151	429
RS6960030	89825608	8	5	1117	0
RS17144825	21677467	9	3	550	137
RS10279821	128683547	10	86	16	20
RS17161157	139521462	18	2	2042	0
RS2699717	118468611	23	4	1940	0
RS11970855	71280299	13	6	13482	0
RS17156813	82515876	20	8	22253	481
RS6979581	108506274	11	9	8521	0
RS10488631	**128594183**	**122**	**28**	**1**	**6**
RS13246321	128701331	0	0	2	7
RS3807306	128580680	0	224	4	2
RS4728142	128573967	106	96	5	3
RS17425212	128721724	0	0	6	546
RS752637	128579420	0	0	7	177
RS3757385	128577304	0	110	8	54
RS17340646	128722514	0	0	9	117
RS10954213	128589427	0	0	10	164
RS6973874	26657873	127	121	25	4
RS4723133	32247043	132	168	384	5
RS10242586	13497326	226	257	127	8
RS40634	81892962	211	364	3014	9
RS865860	42121553	178	303	35	10

Table 60: Chromosome 7 - Top ten SNPs for each method

Chromosome 7				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS10264693	-1.87E+00	-1.55E+00	1.47E-36	1.11E-08
RS12540166	1.44E+00	5.69E-01	6.94E-01	5.39E-01
RS17880470	1.43E+00	3.23E-01	9.17E-02	0
RS17161201	1.27E+00	2.52E-01	1.69E-01	0
RS12535041	-1.20E+00	-4.48E-01	1.61E-06	0
RS10225000	-1.15E+00	-3.35E-01	2.41E-03	0
RS17161202	1.14E+00	3.53E-01	5.56E-01	3.87E-01
RS6960030	-1.14E+00	-5.96E-01	4.66E-03	0
RS17144825	1.09E+00	6.70E-01	1.46E-03	2.00E-01
RS10279821	-1.07E+00	-1.66E-01	2.05E-15	5.26E-02
RS17161157	-8.54E-01	-6.74E-01	1.25E-02	0
RS2699717	-7.07E-01	-6.21E-01	1.15E-02	0
RS11970855	1.00E+00	5.79E-01	2.63E-01	0
RS17156813	7.98E-01	5.14E-01	5.60E-01	4.15E-01
RS6979581	1.04E+00	5.04E-01	1.29E-01	0
RS10488631	**2.97E-01**	**3.27E-01**	**8.86E-43**	**5.85E-03**
RS13246321	0	0	1.28E-42	7.02E-03
RS3807306	0	8.00E-02	2.36E-34	2.79E-04
RS4728142	3.52E-01	1.53E-01	6.29E-32	3.30E-04
RS17425212	0	0	8.36E-27	4.47E-01
RS752637	0	0	1.33E-20	2.32E-01
RS3757385	0	-1.35E-01	2.23E-20	9.83E-02
RS17340646	0	0	5.65E-20	1.81E-01
RS10954213	0	0	2.12E-18	2.28E-01
RS6973874	-1.04E-01	-1.04E-01	3.69E-08	6.60E-04
RS4723133	-1.03E-01	-9.53E-02	7.56E-04	3.80E-03
RS10242586	4.65E-02	7.30E-02	8.77E-05	9.92E-03
RS40634	4.99E-02	5.49E-02	2.42E-02	2.27E-02
RS865860	-5.46E-02	-6.21E-02	1.56E-06	3.28E-02

B.4 Chromosome 8

SNP rs2736340 has been associated with lupus including Gateva et al [60], Bentham et al [4], Marquez et al [107], and Chung et al [61]. This has been reported to be a risk locus for rheumatoid athritis [129], lupus, systemic sclerosis in European populations [130] and was found to be a risk for mucocutaneous lymph node syndrome in Asians. Only the frequentist method ranked it 4th while the others produced a zero-beta coefficient. Meanwhile rs13277113 which is the closest SNP to rs2736340 and has a LD score of $R^2 = 0.978$ and D' = 0.994, with rankings lower than 112th and below across all four methods.

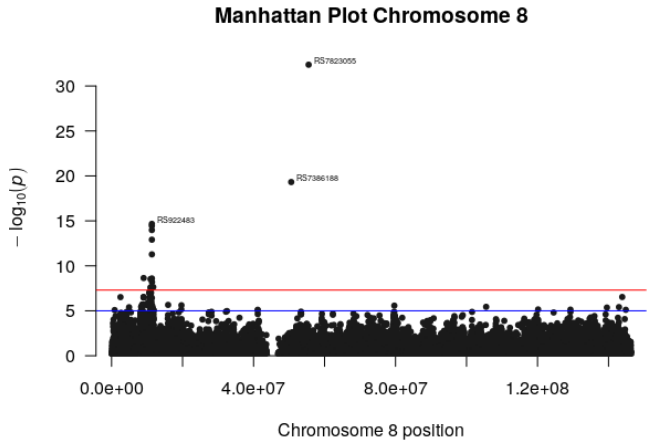

Figure 42: A Manhattan plot of the 3 SNPs with the lowest p-values highlighted

Table 61: Chromosome 8 - Top ten SNPs ranked for each method

Chromosome 8					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS1519371	139549273	1	2	66	0
RS7000460	19803802	2	30	29694	537
RS1470186	19795789	3	979	32901	0
RS10106666	65534695	4	1	7303	0
RS7008834	34567092	5	78	2395	4
RS1882851	91293849	6	0	10526	0
RS7826148	34568015	7	0	1770	34
RS28473203	144873279	8	4	84	1511
RS16889548	118292446	9	3	25102	3
RS16904269	91297472	10	23	7564	1137
RS7823055	55511676	15	5	1	0
RS10090071	108705946	13	6	33193	458
RS17063339	5369835	24	7	32067	118
RS10089147	114878624	12	8	29788	415
RS1358115	4400199	14	9	32861	66
RS16920295	56984164	26	10	11773	126
RS7386188	50609940	0	0	2	0
RS922483	11351912	0	0	3	0
RS2736340	**11343973**	**0**	**0**	**4**	**0**
RS13277113	11349186	112	62	5	31
RS2618476	11352541	0	0	6	0
RS2409781	11359557	0	0	7	0
RS367543	9034148	242	391	8	132
RS1600249	11359638	0	0	9	0
RS7832722	10965442	0	0	10	0
RS4739035	55513201	69	49	12873	1
RS10096780	77402241	128	117	30489	2
RS369240	55523753	111	244	26433	5
RS10101497	77476499	144	214	1209	6
RS2732987	5542087	113	92	21503	7
RS6558668	2426503	114	100	141	8
RS12674710	18263663	45	110	22347	9
RS17243536	122782960	149	237	2419	10

Table 62: Chromosome 8 - Top ten SNPs for each method

Chromosome 8				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS1519371	1.22E+00	6.95E-01	6.71E-06	0
RS7000460	-1.20E+00	-1.94E-01	8.47E-01	7.10E-02
RS1470186	1.11E+00	8.73E-03	9.83E-01	0
RS10106666	-1.10E+00	-9.24E-01	8.63E-02	0
RS7008834	1.04E+00	1.29E-01	1.16E-02	1.55E-07
RS1882851	9.55E-01	0	1.60E-01	0
RS7826148	-9.51E-01	0	6.49E-03	1.68E-04
RS28473203	-9.11E-01	-4.90E-01	9.61E-06	4.22E-01
RS16889548	-8.84E-01	-6.34E-01	6.64E-01	1.25E-07
RS16904269	8.40E-01	2.24E-01	9.12E-02	2.65E-01
RS7823055	7.55E-01	3.89E-01	4.45E-31	0
RS10090071	7.69E-01	3.81E-01	9.95E-01	5.44E-02
RS17063339	6.75E-01	3.64E-01	9.47E-01	4.55E-03
RS10089147	7.77E-01	3.63E-01	8.51E-01	4.51E-02
RS1358115	7.65E-01	3.52E-01	9.81E-01	1.25E-03
RS16920295	6.65E-01	3.37E-01	1.95E-01	5.03E-03
RS7386188	0	0	4.49E-20	0
RS922483	0	0	2.27E-15	0
RS2736340	**0**	**0**	**2.96E-15**	**0**
RS13277113	2.58E-01	1.41E-01	8.99E-15	1.24E-04
RS2618476	0	0	1.08E-13	0
RS2409781	0	0	6.42E-12	0
RS367543	4.30E-02	3.45E-02	1.87E-09	5.60E-03
RS1600249	0	0	2.66E-09	0
RS7832722	0	0	2.85E-09	0
RS4739035	4.50E-01	1.64E-01	2.28E-01	1.99E-08
RS10096780	1.10E-01	8.61E-02	8.79E-01	5.69E-08
RS369240	2.76E-01	5.02E-02	7.18E-01	1.27E-06
RS10101497	6.26E-02	5.53E-02	3.14E-03	1.34E-06
RS2732987	1.26E-01	1.08E-01	5.26E-01	2.02E-06
RS6558668	-1.26E-01	-9.61E-02	2.79E-05	2.17E-06
RS12674710	5.38E-01	9.27E-02	5.57E-01	4.39E-06
RS17243536	5.85E-02	5.14E-02	1.18E-02	4.67E-06

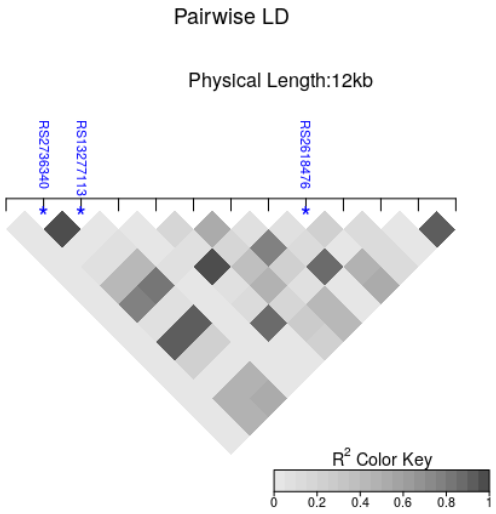

Figure 43: A LD heatmap of a section of gene *BLK*. The three SNPs highlighted are in strong linkage disequilibrium.

Each SNP in Figure 43 has been observed as associated by three different studies (Gateva et al rs2736340 [60], Hom et al [7] rs13277113 and Graham et al [59] rs2618476). These SNPs were ranked 4th,5th and 6th respectively by the frequentist method.

B.5 Chromosome 9

For chromosome 9, from 29764 SNPs, spike method produced 1349 non-zero variables, lasso method produced 2390 and EBEN produced 1549.
Langefeld et al [63] made an association in the gene *AK057451* at SNP rs11788118 in European Americans but no real association was made in this study.

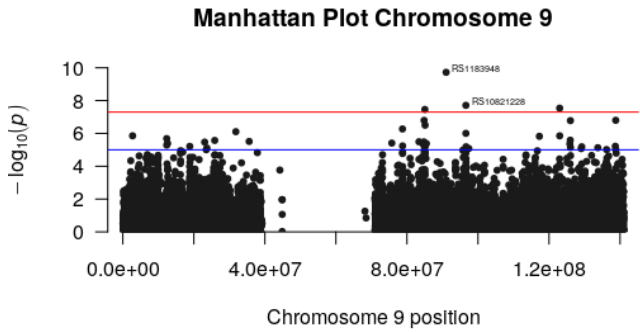

Figure 44: A Manhattan plot with SNPs rs1183948 and rs10821228 highlighted

Table 63: Chromosome 9 - Top ten SNPs ranked for each method

Chromosome 9					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS3750538	90283514	1	15	225	186
RS3750539	90283509	2	115	29529	0
RS1935314	75729671	3	1	26	0
RS563666	136201470	4	6	19121	670
RS7869843	8746568	5	29	14729	0
RS7862733	21632534	6	136	4942	0
RS10975931	6951639	7	2	83	0
RS2065159	132830460	8	27	28006	127
RS7851378	132882000	9	444	22607	0
RS2029647	37223751	10	9	23701	418
RS7031192	129827789	12	7	286	0
RS7033105	138017572	24	3	3538	86
RS12683163	117815011	13	4	285	0
RS7866230	27474551	21	5	4454	190
RS873836	129567615	27	8	11969	0
RS7038413	139084388	36	10	14284	0
RS1183948	91025781	0	103	1	1202
RS10821228	96586062	215	424	2	8
RS2058485	122996420	128	148	3	1
RS951577	85040109	289	889	4	78
RS3119680	138781516	211	628	5	11
RS6478601	126017354	0	1998	6	0
RS2209725	84828677	0	0	7	0
RS1411664	85106906	0	0	8	0
RS10869701	78740727	402	257	9	124
RS7026194	31822916	0	114	10	0
RS10759637	116025024	150	625	16363	2
RS10974623	4560243	133	192	134	3
RS7869617	113060439	152	482	597	4
RS10810534	16233085	144	299	53	5
RS1173099	93441674	157	320	2582	6
RS1061407	6532544	135	315	81	7
RS12339193	76185813	170	219	1023	9
RS7854358	2759058	147	338	14	10

Table 64: Chromosome 9 - Top ten SNPs for each method

Chromosome 9				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS3750538	-2.21E+00	-4.40E-01	2.28E-04	9.64E-02
RS3750539	1.87E+00	1.54E-01	9.89E-01	0
RS1935314	1.29E+00	7.68E-01	6.05E-06	0
RS563666	1.20E+00	5.93E-01	5.26E-01	3.00E-01
RS7869843	1.07E+00	3.44E-01	3.49E-01	0
RS7862733	1.03E+00	1.39E-01	6.05E-02	0
RS10975931	1.01E+00	6.61E-01	3.23E-05	0
RS2065159	9.94E-01	3.63E-01	9.24E-01	6.95E-02
RS7851378	9.67E-01	5.47E-02	6.78E-01	0
RS2029647	9.62E-01	5.10E-01	7.26E-01	2.04E-01
RS7031192	-8.97E-01	-5.57E-01	3.44E-04	0
RS7033105	-8.09E-01	-6.48E-01	3.43E-02	4.51E-02
RS12683163	9.35E-01	6.23E-01	3.43E-04	0
RS7866230	-8.38E-01	-6.13E-01	5.08E-02	9.95E-02
RS873836	7.80E-01	5.14E-01	2.52E-01	0
RS7038413	6.83E-01	4.90E-01	3.33E-01	0
RS1183948	0	-1.67E-01	1.55E-10	5.29E-01
RS10821228	-4.81E-02	-5.72E-02	2.19E-08	5.65E-03
RS2058485	1.31E-01	1.34E-01	2.19E-08	4.28E-04
RS951577	-3.65E-02	-2.95E-02	4.01E-08	4.17E-02
RS3119680	-4.89E-02	-4.16E-02	1.61E-07	7.54E-03
RS6478601	0	5.15E-03	1.91E-07	0
RS2209725	0	0	2.05E-07	0
RS1411664	0	0	3.65E-07	0
RS10869701	2.71E-02	7.74E-02	7.25E-07	6.66E-02
RS7026194	0	-1.54E-01	9.66E-07	0
RS10759637	-6.72E-02	-4.17E-02	4.11E-01	1.01E-03
RS10974623	-8.65E-02	-1.03E-01	9.03E-05	3.33E-03
RS7869617	-6.65E-02	-5.13E-02	1.41E-03	3.77E-03
RS10810534	7.11E-02	7.42E-02	1.69E-05	3.94E-03
RS1173099	6.41E-02	6.99E-02	2.05E-02	4.36E-03
RS1061407	-8.15E-02	-7.15E-02	3.15E-05	4.97E-03
RS12339193	5.94E-02	9.56E-02	3.87E-03	5.83E-03
RS7854358	6.99E-02	6.80E-02	1.54E-06	6.28E-03

B.6 Chromosome 10

For chromosome 10, from 34698 SNPs, spike method produced 1469 non-zero variables, lasso method produced 2413 and EBEN method produced 1551.

An associated SNP rs4948496 from an Asian and European population reported by Morris et al [43] and was observed by Bentham et al [4] was ranked 6th by the frequentist method, although had zero-coefficients with the other methods. The adjacent SNPs are in strong linkage disequilibrium and have been calculated as the SNP of choice by the variable selection methods. SNP rs10821949 was ranked 13th by EBEN method meanwhile SNP rs6479782 has joined SNP rs10821949 and shared the coefficients for spike and lasso methods producing reduced ranks. (See Figure 46 below).

In Bentham et al study, SNP rs2663052 was found to be a risk locus for lupus. The EBEN method produced a ranking of 45th while the frequentist calculated a p-value of 8.72E-05 with a ranking of 187th, spike and slab methods ranked 151st and the lasso ranked 256th.

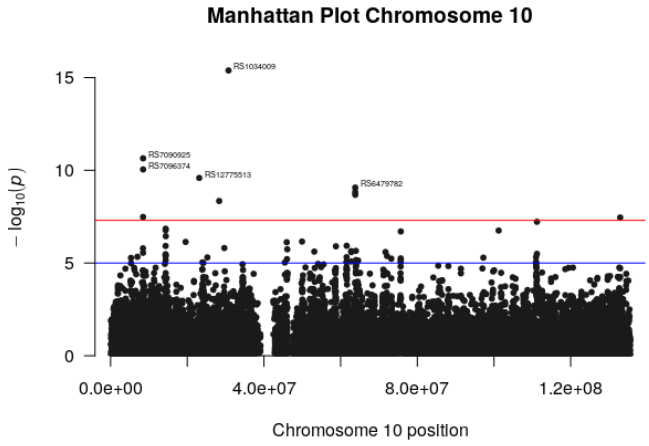

Figure 45: A Manhattan plot of the SNPs with the five lowest p-values highlighted

Table 65: Chromosome 10 - Top ten SNPs ranked for each method

Chromosome 10					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS17100053	84179976	1	1	159	0
RS10128274	31690213	2	27	17175	0
RS10826927	31559619	3	11	21564	105
RS10490914	27633772	4	265	7095	0
RS7078512	27632128	5	269	11095	0
RS7921206	73999284	6	52	11206	0
RS11000229	73993518	7	9	19438	61
RS12253008	84153036	8	14	9972	0
RS12249119	14399180	9	7	10059	43
RS7902030	57796442	10	4	208	0
RS1915446	63063518	11	2	6696	379
RS3765181	18836931	13	3	12424	34
RS4980199	125450986	18	5	340	0
RS4609552	33123586	14	6	603	0
RS12775513	23120547	28	8	4	11
RS17094705	118179627	32	10	22074	993
RS1034009	30757105	0	1238	1	633
RS7090925	8479868	367	570	2	75
RS7096374	8484113	0	0	3	0
RS6479782	63806809	1100	619	5	0
RS4948496	**63805617**	**0**	**0**	**6**	**0**
RS10821949	63811678	149	466	7	13
RS10826385	28326647	26	21	8	24
RS2646425	8470387	304	1115	9	0
RS11017735	132914909	162	122	10	48
RS10509601	92162800	100	77	12518	1
RS7922067	94819606	551	1567	10844	2
RS2441764	57141578	15	179	7579	3
RS3750768	71859747	169	189	32855	4
RS1782741	92226605	142	379	30734	5
RS7916325	57180675	16	715	5506	6
RS4750511	14391273	146	221	16	7
RS1878249	37416376	191	483	28267	8
RS4082517	71704566	145	170	32	9
RS10903888	2909884	152	187	1303	10

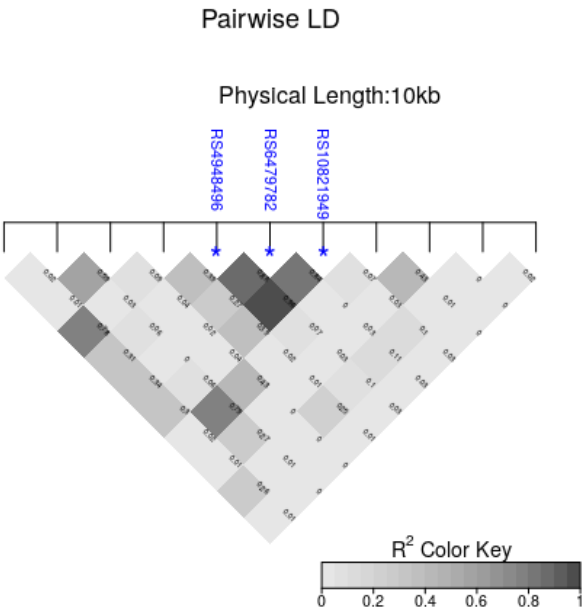

Figure 46: Strong LD between the associated SNP rs4948496 and the closest two SNPs

Table 66: Chromosome 10 - Top ten SNPs for each method

Chromosome 10				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS17100053	-1.73E+00	-1.36E+00	6.56E-05	0
RS10128274	1.43E+00	2.94E-01	3.59E-01	0
RS10826927	1.30E+00	4.49E-01	5.08E-01	3.17E-03
RS10490914	1.16E+00	9.03E-02	8.37E-02	0
RS7078512	1.13E+00	8.99E-02	1.77E-01	0
RS7921206	1.10E+00	2.27E-01	1.80E-01	0
RS11000229	-1.07E+00	-5.03E-01	4.35E-01	1.06E-03
RS12253008	1.05E+00	4.02E-01	1.48E-01	0
RS12249119	-9.60E-01	-5.31E-01	1.50E-01	4.32E-04
RS7902030	9.38E-01	6.33E-01	1.11E-04	0
RS1915446	-9.21E-01	-6.46E-01	7.59E-02	4.74E-02
RS3765181	9.05E-01	6.35E-01	2.14E-01	1.49E-04
RS4980199	7.68E-01	5.45E-01	3.64E-04	0
RS4609552	-8.52E-01	-5.38E-01	1.19E-03	0
RS12775513	-6.70E-01	-5.05E-01	1.98E-10	1.25E-05
RS17094705	6.11E-01	4.53E-01	5.25E-01	1.88E-01
RS1034009	0	1.96E-02	3.63E-16	1.11E-01
RS7090925	3.44E-02	4.95E-02	2.19E-11	1.65E-03
RS7096374	0	0	8.66E-11	0
RS6479782	7.13E-03	4.58E-02	1.02E-09	0
RS4948496	**0**	**0**	**1.47E-09**	**0**
RS10821949	-1.06E-01	-5.82E-02	2.04E-09	1.91E-05
RS10826385	-6.99E-01	-3.38E-01	4.71E-09	7.76E-05
RS2646425	-4.05E-02	-2.25E-02	2.94E-08	0
RS11017735	-8.21E-02	-1.06E-01	4.21E-08	5.08E-04
RS10509601	-4.07E-01	-1.90E-01	2.16E-01	3.96E-11
RS7922067	-2.38E-02	-1.24E-02	1.70E-01	1.59E-10
RS2441764	-8.31E-01	-1.24E-01	9.31E-02	2.36E-09
RS3750768	7.58E-02	1.19E-01	9.30E-01	4.72E-08
RS1782741	-2.86E-01	-7.08E-02	8.48E-01	1.87E-07
RS7916325	-7.75E-01	-3.94E-02	5.51E-02	5.51E-07
RS4750511	-1.19E-01	-1.07E-01	3.07E-07	6.57E-07
RS1878249	-6.48E-02	-5.69E-02	7.55E-01	7.26E-07
RS4082517	-1.23E-01	-1.27E-01	3.67E-06	1.12E-06
RS10903888	9.89E-02	1.19E-01	4.61E-03	2.90E-06

B.7 Chromosome 11

For chromosome 11, from 32336 SNPs, spike method produced 1319 non-zero variables, lasso method produced 1648 and EBEN method produced 788.

SNP rs12802200 [4] ranked 6th by the frequentist method, p-value of 3.05E-11 and featured in the top 125 ranked SNPs for the spike and lasso method.

All 4 methods found the SNP rs7941765 that was reported originally by Bentham et al and followed up by Morris et al [43] and Din et al [131].

Bentham et al reported rs3794060 has an association with lupus and all four methods have produced strong hits. Also reported was SNP rs2732549 has an association with lupus. This was backed up by Langefeld et al [63]. The frequentist method ranked it third while the variable selection methods calculated zero-coefficients.

In a study by Han et al [102] with an Asian population the SNP rs4639966 was associated with lupus but this study had no hits.

The Morris et al study made an association with SNP rs494003 and the frequentist method produced a robust result, but the other three methods achieved zero-coefficients.

SNP rs2732552 in a report by Lessard et al [132] in a wide range of populations produced three zero-coefficients but a significant p-value of 1.81E-10 in the frequentist method.

Another significant result from the frequentist method was for SNP rs4963128 originally reported by Gateva et al [60]. With a p-value of 5.59E-09 and a rank of 25th. It also produced three zero-coefficients.

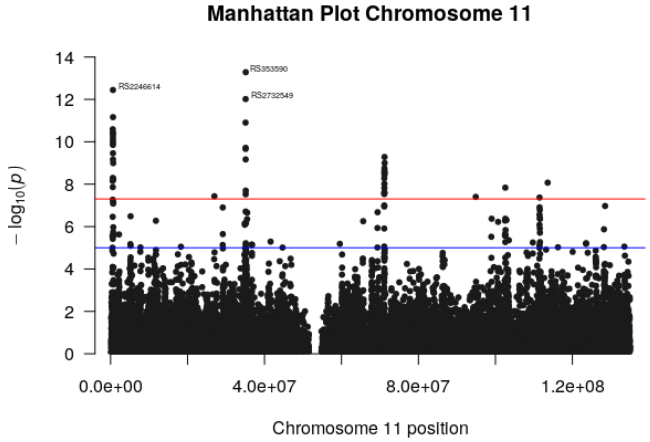

Figure 47: A Manhattan plot of the three SNPs with the lowest p-values highlighted

Table 67: Chromosome 11 - Top ten SNPs ranked for each method

Chromosome 11					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS2730034	62577233	1	1108	8770	0
RS2071035	62623017	2	0	27703	0
RS10790651	124082136	3	30	13962	50
RS4271385	22257284	4	53	9061	0
RS2568030	8926205	5	4	28141	295
RS11026467	22250659	6	11	7537	723
RS11219545	124136439	7	1130	16953	0
RS7123689	33695449	8	55	5579	627
RS1516985	35474901	9	1	61	0
RS12290650	33691318	10	28	5787	82
RS3842724	2185556	17	2	82	0
RS11603987	120920430	14	3	23430	16
RS669724	88712013	35	5	24976	1
RS7123001	130232220	31	6	16618	44
RS679051	68156821	40	7	11116	18
RS12418509	27566176	23	8	24817	55
RS7118237	20299460	28	9	8901	74
RS3136532	46760913	37	10	9705	241
RS353590	35120741	493	264	1	422
RS2246614	619789	0	191	2	9
RS2732549	**35088399**	**0**	**0**	**3**	**0**
RS11246221	630124	0	1501	4	633
RS693163	35074665	1302	481	5	392
RS12802200	**566936**	**124**	**100**	**6**	**314**
RS11246217	623765	0	0	7	0
RS11246213	612967	0	497	8	0
RS10902178	612843	0	0	9	0
RS12805435	612355	0	0	10	0
RS11219741	99262251	92	136	12728	2
RS10431006	49010832	11	17	2353	3
RS593525	65727799	0	0	10012	4
RS1893361	74399425	135	106	3690	5
RS552130	65732800	0	1077	5234	6
RS7932189	44691091	127	161	120	7
RS7118447	99015224	821	1308	3604	8
RS10892549	120075574	131	195	122	10

Table 68: Chromosome 11 - Top ten SNPs for each method

Chromosome 11				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS2730034	1.45E+00	9.41E-03	1.34E-01	0
RS2071035	-1.43E+00	0	8.09E-01	0
RS10790651	1.42E+00	2.57E-01	2.80E-01	2.62E-03
RS4271385	1.34E+00	2.03E-01	1.41E-01	0
RS2568030	1.34E+00	5.28E-01	8.27E-01	4.62E-02
RS11026467	1.29E+00	3.86E-01	1.06E-01	1.92E-01
RS11219545	-1.27E+00	-8.88E-03	3.84E-01	0
RS7123689	-1.23E+00	-1.93E-01	6.33E-02	1.45E-01
RS1516985	1.15E+00	6.73E-01	4.31E-07	0
RS12290650	1.10E+00	2.66E-01	6.76E-02	7.09E-03
RS3842724	9.22E-01	5.79E-01	2.29E-06	0
RS11603987	1.03E+00	5.29E-01	6.37E-01	4.84E-04
RS669724	-6.64E-01	-5.27E-01	6.97E-01	5.14E-09
RS7123001	-6.99E-01	-4.73E-01	3.73E-01	2.38E-03
RS679051	6.35E-01	4.66E-01	1.95E-01	5.93E-04
RS12418509	7.95E-01	4.32E-01	6.91E-01	3.28E-03
RS7118237	7.09E-01	4.28E-01	1.37E-01	5.71E-03
RS3136532	6.59E-01	4.16E-01	1.57E-01	3.57E-02
RS353590	2.29E-02	5.28E-02	5.53E-14	7.63E-02
RS2246614	0	-7.13E-02	4.62E-13	1.96E-04
RS2732549	**0**	**0**	**9.22E-13**	**0**
RS11246221	0	1.68E-03	1.13E-11	1.50E-01
RS693163	3.19E-04	3.13E-02	1.25E-11	6.96E-02
RS12802200	**-2.65E-01**	**-1.11E-01**	**3.05E-11**	**5.11E-02**
RS11246217	0	0	5.06E-11	0
RS11246213	0	3.04E-02	6.24E-11	0
RS10902178	0	0	9.23E-11	0
RS12805435	0	0	1.19E-10	0
RS11219741	-3.79E-01	-9.34E-02	2.42E-01	1.56E-06
RS10431006	-1.08E+00	-3.39E-01	1.49E-02	1.97E-06
RS593525	0	0	1.65E-01	6.37E-05
RS1893361	8.43E-02	1.16E-01	3.10E-02	6.91E-05
RS552130	0	1.01E-02	5.65E-02	1.22E-04
RS7932189	-9.82E-02	-8.15E-02	1.21E-05	1.23E-04
RS7118447	1.08E-02	4.87E-03	2.97E-02	1.33E-04
RS10892549	-8.84E-02	-7.09E-02	1.23E-05	1.97E-04

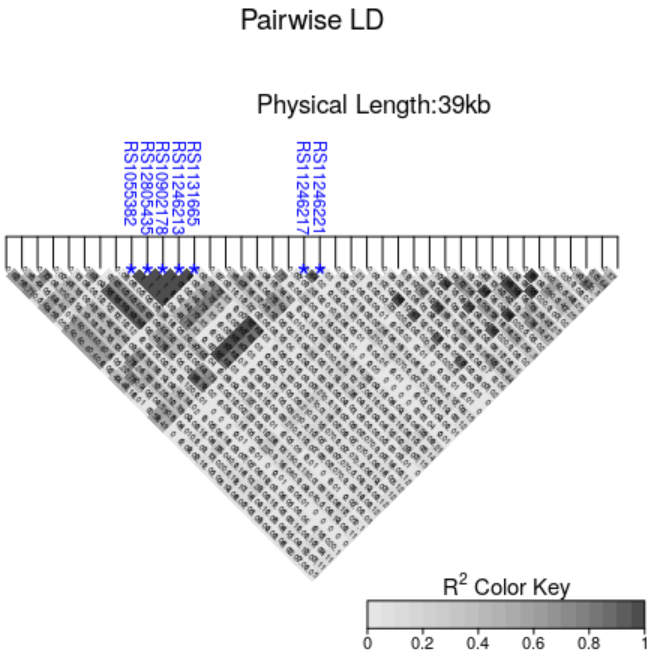

Figure 48: A block of SNPs in near perfect linkage disequilibrium between 5 of the top 10 SNPs by the frequentist method

B.8 Chromosome 12

For chromosome 12, from 31785 SNPs, spike method produced 1280 non-zero variables, lasso method produced 1879 and EBEN method produced 1504.

Bentham et al [4] discovered the association between SNP rs10774625 and lupus, with all four methods finding the hit. It ranked 14th in the frequentist method, 156th in EBEN method and was also found in the other methods but not highly ranked. Morris et al 2016 [43] combined study of European and Chinese populations also found the associated SNP rs10774625.

SNP rs1059312 in the gene *SLC15A4* failed to be found by the spike method although the other three methods did achieve some results. This locus has been found in European, African American, Hispanic, and East Asian populations in Studies by Bentham et al [4], Morris et al, Lessard et al [114] and Langefeld et al [63], respectively.

In the Langefeld et al study of a wide range of populations, SNP rs17005500 in the gene *SYT1* was found to be associated. Spike method did not find this, but the other three methods did.

In the gene *GPR19*, the SNP rs34330 was found by the frequentist method and EBEN although weakly. The study by Yang et al [118] was based on Asian populations.

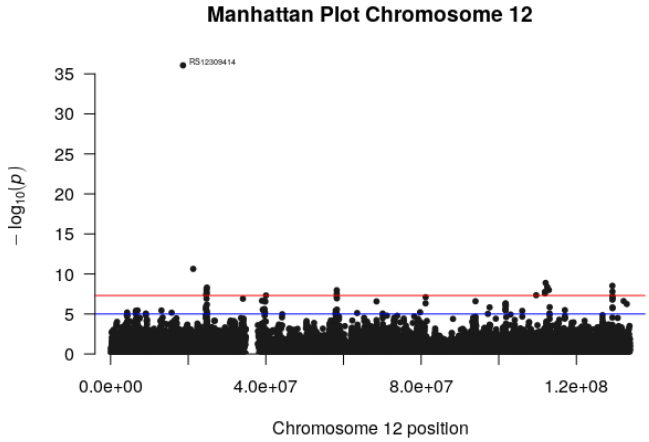

Figure 49: A Manhattan plot with SNP rs12309414 highlighted

Table 69: Chromosome 12 - Top ten SNPs ranked for each method

Chromosome 12					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS12309414	18665717	1	1	1	0
RS1564363	21292864	2	3	2	15
RS1797738	31943705	3	120	4464	0
RS10506079	31942106	4	74	6224	796
RS3741809	68595144	5	62	2940	106
RS7294681	105636626	6	597	6272	0
RS10748101	68595787	7	139	5014	0
RS10844643	33734888	8	51	232	0
RS12303965	79206944	9	14	121	0
RS17014005	86813440	10	148	5287	0
RS784892	53822884	13	2	27122	120
RS11047429	24569816	21	4	1275	0
RS2159889	8932110	14	5	9260	0
RS2114900	30919900	25	6	149	0
RS11052707	33589388	45	7	118	0
RS7971685	16156039	46	8	509	0
RS16934306	29630386	31	9	4638	1268
RS1305267	81165156	26	10	22	0
RS11065987	112072424	0	0	3	977
RS1009858	129261435	0	0	4	0
RS7967537	24819762	480	1222	5	572
RS17696736	112486818	0	0	6	621
RS970290	24730245	0	0	7	0
RS10747786	58260601	0	0	8	115
RS11066320	112906415	211	0	9	1031
RS1487657	24722254	0	0	10	0
RS2364484	6511996	122	98	57	1
RS6539072	103777832	128	160	2025	2
RS2667444	34092598	232	661	23	3
RS10842027	22625661	117	52	2263	4
RS201403	99230509	164	333	127	5
RS11060370	129960343	133	179	424	6
RS7965458	12826806	131	235	159	7
RS10773579	129294881	32	318	21636	8
RS17286243	96088546	132	248	645	9
RS7294479	9473157	134	295	771	10

Table 70: Chromosome 12 - Top ten SNPs for each method

Chromosome 12				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS12309414	-2.54E+00	-2.01E+00	5.70E-37	0
RS1564363	-1.71E+00	-5.82E-01	2.88E-11	1.14E-02
RS1797738	1.27E+00	1.16E-01	4.22E-02	0
RS10506079	1.19E+00	1.58E-01	7.65E-02	4.34E-01
RS3741809	-1.14E+00	-1.75E-01	2.00E-02	7.57E-02
RS7294681	-1.07E+00	-3.23E-02	7.75E-02	0
RS10748101	1.06E+00	1.05E-01	5.14E-02	0
RS10844643	1.04E+00	1.97E-01	1.84E-04	0
RS12303965	9.85E-01	3.35E-01	3.94E-05	0
RS17014005	9.62E-01	9.97E-02	5.67E-02	0
RS784892	9.14E-01	5.91E-01	8.05E-01	8.24E-02
RS11047429	-7.83E-01	-5.79E-01	4.49E-03	0
RS2159889	9.05E-01	5.31E-01	1.47E-01	0
RS2114900	-7.63E-01	-5.14E-01	6.49E-05	0
RS11052707	5.67E-01	4.45E-01	3.63E-05	0
RS7971685	-5.68E-01	-4.42E-01	8.03E-04	0
RS16934306	-6.72E-01	-4.16E-01	4.52E-02	6.57E-01
RS1305267	7.51E-01	4.07E-01	1.06E-07	0
RS11065987	0	0	2.19E-09	5.14E-01
RS1009858	0	0	2.96E-09	0
RS7967537	2.28E-02	1.04E-02	5.01E-09	3.25E-01
RS17696736	0	0	6.29E-09	3.54E-01
RS970290	0	0	8.05E-09	0
RS10747786	0	0	1.37E-08	7.95E-02
RS11066320	4.67E-02	0	1.39E-08	5.43E-01
RS1487657	0	0	1.53E-08	0
RS2364484	-1.58E-01	-1.31E-01	4.10E-06	2.45E-05
RS6539072	-1.04E-01	-9.45E-02	1.04E-02	6.79E-04
RS2667444	4.29E-02	2.89E-02	1.11E-07	2.34E-03
RS10842027	-2.70E-01	-1.89E-01	1.24E-02	2.44E-03
RS201403	-5.72E-02	-5.35E-02	4.31E-05	2.61E-03
RS11060370	-8.23E-02	-8.89E-02	5.84E-04	4.84E-03
RS7965458	-9.61E-02	-7.06E-02	7.72E-05	5.25E-03
RS10773579	-6.49E-01	-5.50E-02	5.80E-01	5.35E-03
RS17286243	9.37E-02	6.75E-02	1.15E-03	6.19E-03
RS7294479	8.08E-02	5.91E-02	1.75E-03	6.46E-03

B.9 Chromosome 13

For chromosome 13, from 24265 SNPs, spike method produced 1175 non-zero variables, lasso method produced 1470 and EBEN method produced 1334. There are no recorded SNPs that are associated with lupus on chromosome 13.

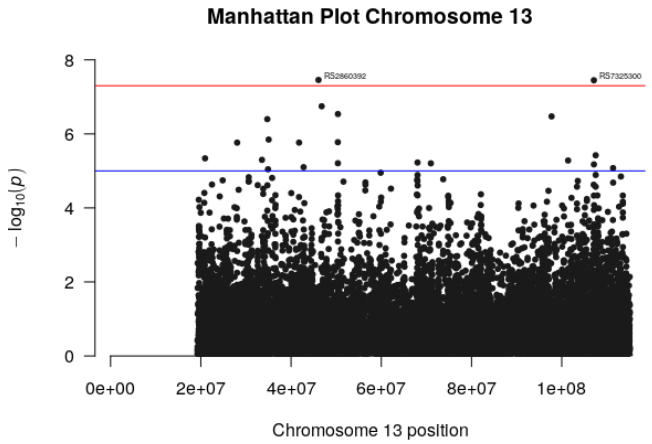

Figure 50: A Manhattan plot with SNP rs2860392 and rs7325300 highlighted

Table 71: Chromosome 13 - Top ten SNPs ranked for each method

Chromosome 13					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS9582104	97738934	1	1	4	0
RS12431281	108845919	2	2	6003	4
RS16972194	108920961	3	5	1334	0
RS2860392	46085180	4	177	2	68
RS2985970	46085864	5	0	21242	0
RS9555736	111497851	6	10	19942	5
RS9529059	66991537	7	667	7982	528
RS9522500	89956070	8	34	2741	0
RS7982531	73547569	9	202	6746	0
RS1373496	66987127	10	0	22647	0
RS8001449	35035782	29	3	7	0
RS7994107	78574175	14	4	22522	469
RS17089102	54521057	33	6	10847	94
RS3013348	56004322	22	7	4192	56
RS9588246	111372460	15	8	23	0
RS865296	108537405	19	9	444	0
RS7325300	107114450	130	181	1	23
RS912784	46785521	126	57	3	206
RS1198329	50390096	505	0	5	0
RS4943189	34736943	0	1452	6	60
RS7992673	41776362	487	431	8	186
RS9535343	50370004	183	222	9	59
RS12429751	28054171	0	212	10	0
RS9574551	36347006	145	189	359	1
RS9598029	34297679	50	15	20764	2
RS9518442	102208507	132	352	4523	3
RS17068766	47314896	66	22	11549	6
RS1183680	112057748	165	300	10951	7
RS6561393	48194319	126	67	21117	8
RS7332131	19619863	136	266	492	9
RS11840971	51615784	134	173	38	10

Table 72: Chromosome 13 - Top ten SNPs for each method

Chromosome 13				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS9582104	-1.67E+00	-1.29E+00	3.22E-07	0
RS12431281	1.32E+00	6.95E-01	1.10E-01	6.64E-04
RS16972194	1.27E+00	5.79E-01	9.19E-03	0
RS2860392	1.19E+00	7.69E-02	3.92E-08	2.34E-02
RS2985970	1.11E+00	0	8.28E-01	0
RS9555736	-1.09E+00	-4.25E-01	7.58E-01	8.57E-04
RS9529059	1.08E+00	2.11E-02	2.11E-02	2.50E-01
RS9522500	1.07E+00	2.20E-01	2.20E-01	0
RS7982531	1.06E+00	6.75E-02	1.32E-01	0
RS1373496	-1.05E+00	0	9.08E-01	0
RS8001449	-7.96E-01	-5.98E-01	1.38E-06	0
RS7994107	9.48E-01	5.86E-01	9.00E-01	2.22E-01
RS17089102	7.51E-01	5.18E-01	2.92E-01	3.79E-02
RS3013348	8.75E-01	5.04E-01	6.04E-02	1.77E-02
RS9588246	9.18E-01	4.52E-01	1.24E-05	0
RS865296	8.84E-01	4.25E-01	1.42E-03	0
RS7325300	6.09E-02	7.52E-02	3.73E-08	4.86E-03
RS912784	-2.88E-01	-1.71E-01	2.10E-07	9.91E-02
RS1198329	-2.12E-02	0	3.53E-07	0
RS4943189	0	-2.15E-04	5.20E-07	5.37E-01
RS7992673	2.20E-02	3.41E-02	1.60E-06	8.91E-02
RS9535343	-5.27E-02	-6.23E-02	1.89E-06	1.85E-02
RS12429751	0	-6.51E-02	2.16E-06	0
RS9574551	-7.00E-02	-7.01E-02	9.47E-04	1.93E-05
RS9598029	5.85E-01	3.70E-01	8.02E-01	5.40E-04
RS9518442	8.23E-02	4.10E-02	6.78E-02	5.97E-04
RS17068766	4.95E-01	3.15E-01	3.28E-01	1.21E-03
RS1183680	-5.91E-02	-4.78E-02	2.97E-01	1.72E-03
RS6561393	2.89E-01	1.64E-01	8.21E-01	1.81E-03
RS7332131	7.55E-02	5.23E-02	1.70E-03	2.01E-03
RS11840971	-8.09E-02	-7.74E-02	2.39E-05	2.01E-03

B.10 Chromosome 14

For chromosome 14, from 20959 SNPs, spike method produced 1438 non-zero variables, lasso method produced 1398 and EBEN method produced 1038.

All four methods found Bentham et al [4] associated SNP rs4902562 but none were highly ranked.

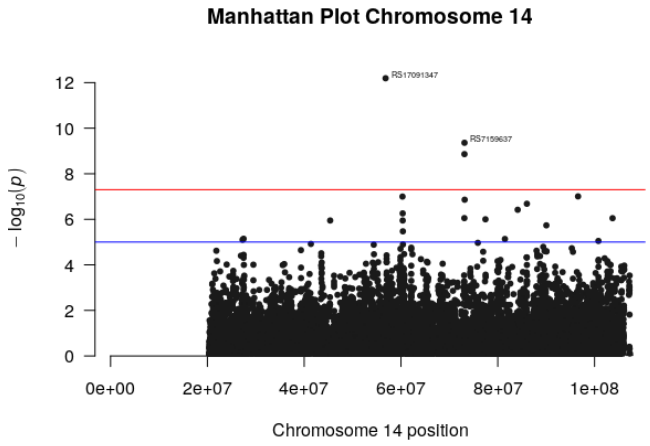

Figure 51: A Manhattan plot with SNPs rs7159637 and rs17091347highlighted

Table 73: Chromosome 14 - Top ten SNPs ranked for each method

Chromosome 14					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS7154718	64143825	1	0	9625	603
RS7142737	64120206	2	1249	57	201
RS17123116	51371081	3	8	408	7
RS10162571	79958705	4	2	61	568
RS8016080	45373245	5	3	13	0
RS11844619	80082435	6	4	11840	47
RS10132507	31032594	7	5	12354	34
RS10139102	59247190	8	27	6274	22
RS7140746	98244968	9	21	2151	0
RS2275466	51372315	10	735	2187	747
RS17091347	56820049	15	1	1	101
RS10130000	35068168	48	6	2659	172
RS10139139	84114680	42	7	9	0
RS10132319	21423801	45	9	129	11
RS6575958	103716196	13	10	10	9
RS7159637	73123376	0	1074	2	0
RS2803977	73114476	55	22	3	392
RS1957309	60312064	306	477	4	709
RS10148669	96584980	64	16	5	721
RS7159986	73162115	487	116	6	165
RS10137902	86013442	0	0	7	0
RS17096723	60352299	0	0	8	0
RS10148260	22926322	18	35	18556	1
RS4906205	102932447	49	317	12989	2
RS12147516	22915816	22	73	8137	3
RS11851013	66016212	17	48	2164	4
RS12431702	30088138	289	386	12650	5
RS1255720	64009521	129	259	9751	6
RS3818263	92588002	228	417	8369	8
RS2096023	RS2096023	124	254	5129	10

Table 74: Chromosome 14 - Top ten SNPs for each method

Chromosome 14				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS7154718	-1.77E+00	0	3.03E-01	1.06E-01
RS7142737	-1.71E+00	-1.68E-03	6.38E-05	1.12E-02
RS17123116	1.42E+00	4.66E-01	1.26E-03	2.61E-08
RS10162571	1.28E+00	7.89E-01	6.75E-05	9.39E-02
RS8016080	1.27E+00	7.59E-01	1.10E-06	0
RS11844619	1.24E+00	6.62E-01	4.25E-01	1.38E-04
RS10132507	1.21E+00	6.20E-01	4.54E-01	4.67E-05
RS10139102	-1.13E+00	-2.66E-01	1.48E-01	4.29E-06
RS7140746	-1.13E+00	-2.95E-01	2.19E-02	0
RS2275466	1.13E+00	1.64E-02	2.26E-02	1.57E-01
RS17091347	1.06E+00	9.71E-01	9.84E-13	1.64E-03
RS10130000	-6.44E-01	-5.09E-01	3.25E-02	7.12E-03
RS10139139	7.09E-01	4.72E-01	5.60E-07	0
RS10132319	-6.68E-01	-4.64E-01	2.26E-04	1.89E-07
RS6575958	-1.10E+00	-4.49E-01	8.86E-07	5.32E-08
RS7159637	0	-6.00E-03	6.93E-10	0
RS2803977	-5.83E-01	-2.89E-01	2.02E-09	4.78E-02
RS1957309	-3.67E-02	-2.88E-02	8.55E-08	1.47E-01
RS10148669	5.35E-01	3.44E-01	1.09E-07	1.51E-01
RS7159986	-2.37E-02	-9.95E-02	1.60E-07	6.57E-03
RS10137902	0	0	2.82E-07	0
RS17096723	0	0	4.84E-07	0
RS10148260	1.04E+00	2.39E-01	8.45E-01	2.40E-14
RS4906205	6.36E-01	4.42E-02	4.91E-01	9.59E-11
RS12147516	9.71E-01	1.52E-01	2.31E-01	3.05E-10
RS11851013	1.04E+00	1.94E-01	2.22E-02	3.66E-09
RS12431702	-3.87E-02	-3.56E-02	4.72E-01	1.48E-08
RS1255720	8.28E-02		3.10E-01	2.32E-08
RS3818263	4.55E-02	3.29E-02	2.42E-01	3.06E-08
RS2096023	8.89E-02	5.35E-02	1.05E-01	1.82E-07

B.11 Chromosome 15

For chromosome 15, from 19521 SNPs, spike method produced 1143 non-zero variables, lasso method produced 1196 and EBEN method produced 747.

All four methods found rs2289583 in the gene *CSK* and were in the top 100 including EBEN method being highest ranked SNP.

All 4 methods found rs8035957 that was picked out in a study by Wen et al [133] in a Chinese population

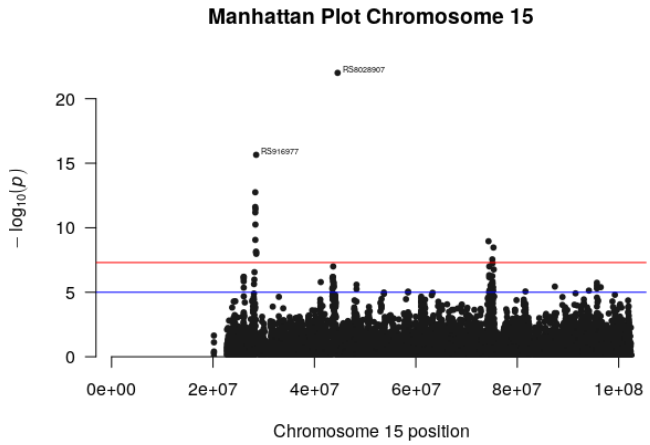

Figure 52: A Manhattan plot with SNPs rs8028907 and rs916977 highlighted

Table 75: Chromosome 15 - Top ten SNPs ranked for each method

Chromosome 15					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS4774617	52553744	1	8	2359	0
RS4496072	52552535	2	4	3060	624
RS7175376	72842192	3	7	5010	678
RS311917	69358613	4	0	1763	0
RS16957064	72827051	5	28	13548	0
RS2470104	48334867	6	1	921	541
RS2290622	51594669	7	12	13096	0
RS311893	69325581	8	0	2007	0
RS8043143	42948674	9	11	10472	0
RS12592573	43035004	10	62	8185	0
RS9920741	69233869	12	2	15609	0
RS16961587	49052809	11	3	13085	0
RS724099	56650343	19	5	899	0
RS2899058	42818026	18	6	18648	0
RS7342673	52799364	40	9	12235	0
RS7178949	80778153	26	10	3850	0
RS8028907	44574649	0	0	1	0
RS916977	28513364	353	299	2	9
RS4778138	28335820	0	0	3	100
RS7174027	28328765	115	287	4	115
RS4778241	28338713	0	0	5	116
RS3935591	28374012	0	0	6	609
RS12593929	28359258	142	131	7	441
RS12917449	74331659	180	142	8	37
RS2289583	**75311036**	**93**	**66**	**9**	**1**
RS7495174	28344238	0	0	10	0
RS925480	70737052	92	34	883	2
RS1947057	86552956	99	78	481	3
RS2656065	78750549	125	205	14379	4
RS17423970	48302064	206	245	56	5
RS4622471	81634574	196	247	210	6
RS16962243	49562732	124	171	496	7
RS2663907	81373032	106	151	160	8
RS627101	64900526	189	301	607	10

Table 76: Chromosome 15 - Top ten SNPs for each method

Chromosome 15				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS4774617	1.50E+00	3.02E-01	3.04E-02	0
RS4496072	1.36E+00	3.85E-01	4.97E-02	6.61E-01
RS7175376	1.26E+00	3.07E-01	1.18E-01	7.12E-01
RS311917	1.19E+00	0	1.84E-02	0
RS16957064	-1.13E+00	-2.13E-01	5.99E-01	0
RS2470104	1.08E+00	6.41E-01	5.09E-03	5.95E-01
RS2290622	1.05E+00	2.92E-01	5.69E-01	0
RS311893	9.96E-01	0	2.33E-02	0
RS8043143	9.88E-01	2.94E-01	4.08E-01	0
RS12592573	9.77E-01	1.42E-01	2.74E-01	0
RS9920741	8.81E-01	5.25E-01	7.38E-01	0
RS16961587	-9.18E-01	-4.65E-01	5.68E-01	0
RS724099	7.70E-01	3.75E-01	4.87E-03	0
RS2899058	7.78E-01	3.48E-01	9.43E-01	0
RS7342673	5.63E-01	3.02E-01	5.13E-01	0
RS7178949	-6.90E-01	-3.00E-01	7.67E-02	0
RS8028907	0	0	2.10E-22	0
RS916977	2.78E-02	3.77E-02	2.04E-15	4.13E-02
RS4778138	0	0	5.12E-13	1.82E-01
RS7174027	7.45E-02	3.91E-02	5.40E-12	2.95E-01
RS4778241	0	0	1.42E-11	1.94E-01
RS3935591	0	0	1.44E-11	6.49E-01
RS12593929	-5.52E-02	-7.65E-02	1.46E-10	4.97E-01
RS12917449	4.85E-02	7.04E-02	7.67E-02	9.61E-02
RS2289583	**1.64E-01**	**1.38E-01**	**2.36E-09**	**9.18E-04**
RS7495174	0	0	2.62E-09	0
RS925480	2.36E-01	1.95E-01	4.68E-03	8.36E-03
RS1947057	1.03E-01	1.19E-01	1.41E-03	9.64E-03
RS2656065	5.96E-02	5.15E-02	5.96E-02	2.26E-02
RS17423970	-4.28E-02	-4.40E-02	6.31E-06	2.86E-02
RS4622471	4.51E-02	4.36E-02	1.71E-04	3.77E-02
RS16962243	5.96E-02	5.91E-02	1.51E-03	3.94E-02
RS2663907	8.28E-02	6.74E-02	9.22E-05	3.97E-02
RS627101	-4.65E-02	-3.76E-02	2.26E-03	4.55E-02

B.12 Chromosome 17

For chromosome 17, from 18084 SNPs, spike method produced 1438 non-zero variables, lasso method produced 1307 and the EBEN method produced 717.

Bentham et al [4], Morris et al [43], and Din et al [131] all found an association with SNP rs2941509 (in the gene *IKZF3*) and lupus but the variable selection methods found no associations (frequentist method was ranked 36th). The marker rs9904834 has $R^2 = 0.971$ and D' = 0.997 in LD with SNP rs2941509. This locus produced a ranking of 75th by spike, 35th by lasso and 133rd by EBEN.

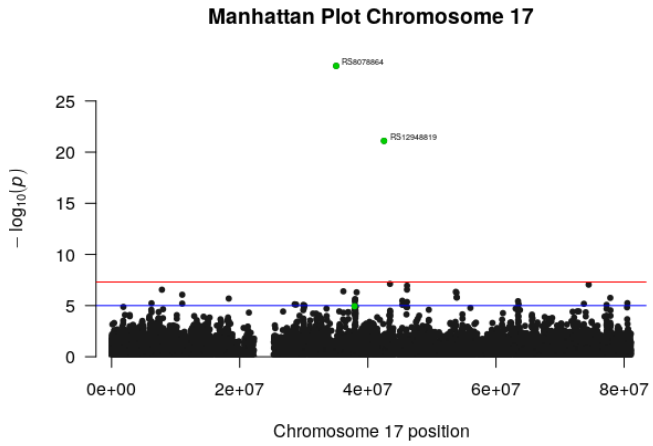

Figure 53: A Manhattan plot with SNPs rs8078864 and rs12948819 highlighted

Table 77: Chromosome 17 - Top ten SNPs ranked for each method

Chromosome 17					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS6503802	55344646	1	338	7970	0
RS12948819	42521824	2	1	2	1
RS16958130	55342130	3	0	7018	0
RS9904838	54046321	4	0	1090	0
RS8079652	58679312	5	10	2174	0
RS9903864	3337960	6	91	283	0
RS9914712	54046852	7	156	1239	0
RS4796054	33536506	8	7	419	0
RS8078864	35056109	9	2	1	0
RS12103525	53906248	10	88	15632	41
RS7209106	30986185	23	3	520	0
RS7226346	544747	19	4	8227	0
RS7212285	58790082	11	5	8514	111
RS2290509	18145450	26	6	13727	236
RS9907400	53733289	29	8	10	0
RS1266474	2534710	27	9	17366	318
RS10468514	43456240	109	166	3	3
RS11553545	74470194	0	67	4	0
RS10491182	46120767	82	46	5	204
RS7209072	46110469	0	0	6	247
RS9897629	7857350	0	58	7	0
RS11653037	36179647	0	0	8	0
RS9905070	53833334	0	0	9	0
RS16970025	47243622	103	169	1600	2
RS4793900	55790197	102	175	2027	4
RS7224279	31450715	85	83	3897	5
RS2074190	45811210	100	176	4503	6
RS2157839	43151400	129	276	5853	7
RS584300	18288199	101	160	15	8
RS4789986	77266813	369	350	86	9
RS10445387	31879590	93	222	253	10

Table 78: Chromosome 17 - Top ten SNPs for each method

Chromosome 17				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS6503802	-1.80E+00	-4.07E-02	2.95E-01	0
RS12948819	-1.74E+00	-1.38E+00	6.84E-22	2.00E-15
RS16958130	1.67E+00	0	2.40E-01	0
RS9904838	-1.53E+00	0	1.01E-02	0
RS8079652	1.45E+00	4.64E-01	3.12E-02	0
RS9903864	1.44E+00	1.10E-01	1.06E-03	0
RS9914712	1.42E+00	6.51E-02	1.06E-03	0
RS4796054	-1.40E+00	-5.36E-01	1.26E-02	0
RS8078864	1.36E+00	1.13E+00	2.05E-03	0
RS12103525	-1.34E+00	-1.12E-01	8.14E-01	7.69E-02
RS7209106	-9.20E-01	-6.58E-01	3.03E-03	0
RS7226346	9.53E-01	6.07E-01	3.10E-01	0
RS7212285	1.31E+00	5.96E-01	3.27E-01	1.59E-01
RS2290509	8.25E-01	5.39E-01	6.75E-01	2.96E-01
RS9907400	-8.16E-01	-5.22E-01	5.97E-07	0
RS1266474	8.31E-01	5.00E-01	9.43E-01	3.80E-01
RS10468514	-5.92E-02	-7.15E-02	7.92E-08	9.86E-03
RS11553545	0	1.44E-01	8.68E-08	0
RS10491182	1.51E-01	1.92E-01	1.01E-07	2.64E-01
RS7209072	0	0	2.50E-07	3.07E-01
RS9897629	0	1.59E-01	3.11E-07	0
RS11653037	0	0	4.12E-07	0
RS9905070	0	0	5.26E-07	0
RS16970025	-8.51E-02	-7.12E-02	1.90E-02	8.80E-03
RS4793900	8.65E-02	6.90E-02	2.77E-02	1.35E-02
RS7224279	1.43E-01	1.18E-01	8.81E-02	1.68E-02
RS2074190	-8.72E-02	-6.89E-02	1.15E-01	1.75E-02
RS2157839	7.19E-02	5.03E-02	1.78E-01	2.37E-02
RS584300	-8.65E-02	-7.67E-02	1.92E-06	2.57E-02
RS4789986	2.38E-02	3.87E-02	9.30E-05	2.67E-02
RS10445387	9.92E-02	5.97E-02	8.82E-04	2.97E-02

B.13 Chromosome 18

For chromosome 18, from 19037 SNPs, spike method produced 1422 non-zero variables, lasso method produced 2386 and EBEN method produced 1037.

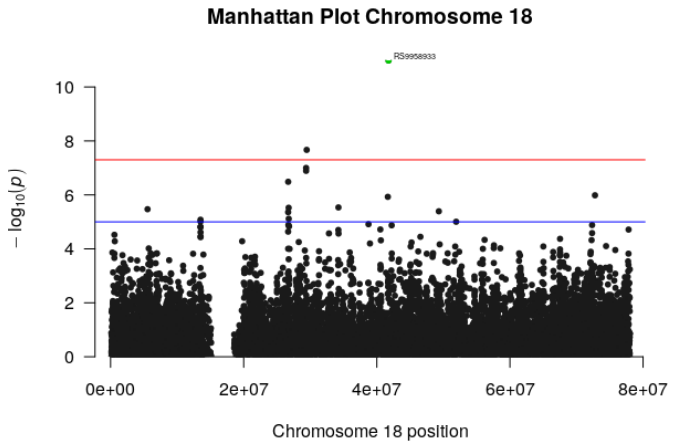

Figure 54: A Manhattan plot with the top ranked SNP rs9958933 by the frequentist method highlighted

Table 79: Chromosome 18 - Top ten SNPs ranked for each method

Chromosome 18					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS9789108	10703865	1	20	12149	0
RS9789103	10703497	2	11	1019	0
RS17088883	71828924	3	34	2589	412
RS9958933	41759986	4	2	1	278
RS17078922	66004354	5	52	18586	96
RS8092546	66005993	6	99	12979	0
RS7226800	72777472	7	1	6	0
RS5022001	10979948	8	31	2848	0
RS7240739	10975049	9	92	10273	0
RS17080367	66774452	10	77	656	0
RS16967782	33929428	13	3	8020	139
RS4798212	4310797	11	4	15399	145
RS17188179	22487270	23	5	7020	289
RS9965651	62253299	22	6	2056	106
RS33969048	21057210	20	7	2100	0
RS12457549	59380988	24	8	10269	11
RS11663711	4286351	17	9	757	0
RS675604	66290862	28	10	753	0
RS1915	29463735	161	176	2	567
RS7226458	29375892	217	1735	3	563
RS10502578	29375087	0	0	4	531
RS11083322	26684465	0	0	5	78
RS1516786	41653101	279	429	7	37
RS17680285	34233545	0	869	8	0
RS1719961	5553097	0	1544	9	153
RS2056095	26753739	781	0	10	307
RS9947927	24761232	52	143	4674	1
RS7238082	42277189	85	43	18902	2
RS17688362	41745680	96	156	11014	3
RS7243961	69432511	90	69	2305	4
RS17077479	65241380	16	66	3608	5
RS10502819	41290085	94	83	4521	6
RS13381189	631211	132	290	40	7
RS592209	48424454	98	537	1915	8
RS2863264	53948968	124	241	516	9
RS1943227	58058711	95	142	610	10

Table 80: Chromosome 18 - Top ten SNPs for each method

Chromosome 18				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS9789108	3.23E+00	4.46E-01	5.38E-01	0
RS9789103	-3.02E+00	-5.42E-01	9.42E-03	0
RS17088883	1.46E+00	3.51E-01	4.46E-02	2.42E-01
RS9958933	-1.44E+00	-1.09E+00	1.20E-11	1.64E-01
RS17078922	1.36E+00	2.94E-01	9.69E-01	6.76E-02
RS8092546	-1.29E+00	-1.99E-01	5.91E-01	0
RS7226800	-1.26E+00	-1.11E+00	9.94E-07	0
RS5022001	-1.20E+00	-3.68E-01	5.29E-02	0
RS7240739	1.05E+00	2.03E-01	4.15E-01	0
RS17080367	1.05E+00	2.39E-01	4.81E-03	0
RS16967782	1.02E+00	7.80E-01	2.79E-01	9.07E-02
RS4798212	1.04E+00	6.78E-01	7.49E-01	9.49E-02
RS17188179	8.32E-01	6.74E-01	2.27E-01	1.70E-01
RS9965651	8.36E-01	6.69E-01	2.99E-02	7.44E-02
RS33969048	-8.64E-01	-6.67E-01	3.10E-02	0
RS12457549	8.14E-01	6.18E-01	4.15E-01	3.94E-03
RS11663711	-9.14E-01	-5.92E-01	5.79E-03	0
RS675604	7.41E-01	5.48E-01	5.70E-03	0
RS1915	-6.36E-02	-1.30E-01	2.37E-08	3.36E-01
RS7226458	5.16E-02	1.01E-02	8.74E-08	3.32E-01
RS10502578	0	0	1.12E-07	3.13E-01
RS11083322	0	0	3.30E-07	5.51E-02
RS1516786	4.42E-02	6.56E-02	1.42E-06	2.73E-02
RS17680285	0	-3.50E-02	2.74E-06	0
RS1719961	0	1.41E-02	3.31E-06	1.01E-01
RS2056095	-1.60E-02	0	3.87E-06	1.78E-01
RS9947927	-5.39E-01	-1.52E-01	1.21E-01	1.18E-04
RS7238082	3.48E-01	3.16E-01	9.91E-01	6.46E-04
RS17688362	-1.17E-01	-1.40E-01	4.63E-01	9.33E-04
RS7243961	-1.83E-01	-2.57E-01	3.61E-02	1.43E-03
RS17077479	9.23E-01	2.65E-01	7.87E-02	1.62E-03
RS10502819	2.00E-01	2.32E-01	1.15E-01	2.08E-03
RS13381189	7.28E-02	8.96E-02	5.70E-05	2.30E-03
RS592209	1.16E-01	4.78E-02	2.67E-02	2.39E-03
RS2863264	-7.54E-02	-1.01E-01	3.13E-03	3.25E-03
RS1943227	-1.21E-01	-1.52E-01	4.19E-03	3.75E-03

B.14 Chromosome 19

For chromosome 19, from 13157 SNPs, spike method produced 1181 non-zero variables, lasso method produced 3345 and EBEN method produced 990.

The top ranked SNP by the frequentist and EBEN methods, rs2304256 in the *TYK2* gene has been reported to show evidence of association with SLE in a Finnish case-control study [134], whereas a Japanese case-control study found no association [135]. Bentham et al [4] and Morris et al [43] both found associations with SLE and the SNP.

In a study by Kim et al [136] in European, African, Hispanic and Korean populations found the SNP rs3093030. The frequentist method produced a p-value of 5.36E-08 and was ranked 8th. The variable selection methods produced zero-coefficients.

Langefeld et al [63] in 2017 found an association with SNP rs13344313 and SLE. The frequentist method made it the 136th top ranked SNP while the variable selections failed to note it.

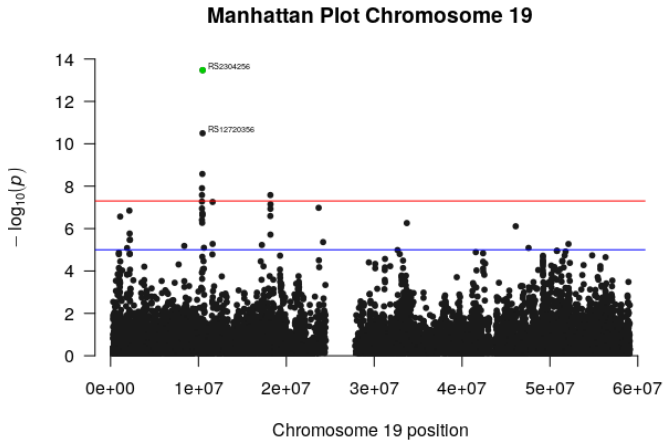

Figure 55: A Manhattan plot with SNPs rs2304256 and rs12720356 highlighted

Table 81: Chromosome 19 - Top ten SNPs ranked for each method

Chromosome 19					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS8109273	12224172	1	23	1487	0
RS10423417	12221700	2	21	1430	0
RS16970737	36368048	3	5	6193	0
RS457155	36356569	4	22	2917	0
RS8110546	52383556	5	47	6250	0
RS8110461	52383569	6	56	6692	502
RS9304831	32266516	7	29	5661	0
RS8105148	16717675	8	3	1793	0
RS17722868	16301100	9	67	7978	41
RS751403	16311070	10	188	8104	0
RS12460179	7742284	24	1	78	0
RS10403531	23680857	13	2	10	0
RS2650822	21651310	69	4	541	0
RS10407428	31821763	34	6	75	0
RS10409643	44681875	28	7	404	0
RS16972885	39421751	54	8	6235	0
RS17725970	44387820	26	9	1881	187
RS2683028	15774241	31	10	10350	366
RS2304256	**10475652**	**65**	**942**	**1**	**1**
RS12720356	10469975	0	732	2	251
RS2278442	10444826	79	198	3	6
RS2228615	10403368	799	2059	4	279
RS2569693	10399904	0	0	5	625
RS447009	18185192	78	164	6	48
RS624899	11612498	84	215	7	151
RS3093030	10397403	0	0	8	0
RS436857	18197635	0	0	9	0
RS7254835	21390608	71	473	118	2
RS11083430	38139488	26	138	496	3
RS10427026	32685150	72	114	302	4
RS7250471	56898853	77	152	1094	5
RS757228	1101992	81	537	58	7
RS4806860	945710	75	246	84	8
RS2972515	48464978	110	1158	211	9
RS601338	49206674	82	654	50	10

Table 82: Chromosome 19 - Top ten SNPs for each method

Chromosome 19				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS8109273	2.08E+00	6.14E-01	2.86E-02	0
RS10423417	1.96E+00	6.22E-01	2.70E-02	0
RS16970737	-1.83E+00	-9.67E-01	3.29E-01	0
RS457155	1.74E+00	6.16E-01	9.05E-02	0
RS8110546	1.68E+00	4.19E-01	3.34E-01	0
RS8110461	1.68E+00	3.76E-01	3.72E-01	3.88E-01
RS9304831	-1.34E+00	-5.44E-01	2.83E-01	0
RS8105148	-1.32E+00	-1.01E+00	3.85E-02	0
RS17722868	1.24E+00	3.59E-01	4.89E-01	4.13E-02
RS751403	1.22E+00	1.90E-01	5.00E-01	0
RS12460179	-9.44E-01	-1.10E+00	6.18E-05	0
RS10403531	1.13E+00	1.06E+00	1.01E-07	0
RS2650822	2.86E-01	9.70E-01	4.26E-03	0
RS10407428	7.94E-01	9.58E-01	5.83E-05	0
RS10409643	-8.50E-01	-8.09E-01	2.57E-03	0
RS16972885	-4.52E-01	-8.04E-01	3.33E-01	0
RS17725970	8.93E-01	7.98E-01	4.17E-02	1.60E-01
RS2683028	-8.28E-01	-7.78E-01	7.14E-01	2.87E-01
RS2304256	**-1.43E-01**	**-6.52E-02**	**2.86E-14**	**9.65E-05**
RS12720356	0	-7.76E-02	3.48E-11	1.99E-01
RS2278442	1.02E-01	1.82E-01	2.72E-09	2.69E-03
RS2228615	1.08E-02	2.49E-02	1.40E-08	2.19E-01
RS2569693	0	0	2.84E-08	4.66E-01
RS447009	-1.05E-01	-1.42E-01	3.00E-08	5.13E-02
RS624899	9.09E-02	1.69E-01	4.81E-08	1.32E-01
RS3093030	0	0	5.36E-08	0
RS436857	0	0	6.92E-08	0
RS7254835	1.34E-01	1.06E-01	1.68E-04	6.24E-04
RS11083430	-1.24E-01	-2.29E-01	3.68E-03	8.39E-04
RS10427026	1.33E-01	2.58E-01	1.39E-03	1.00E-03
RS7250471	1.07E-01	2.18E-01	1.66E-02	2.05E-03
RS757228	-9.55E-02	-9.78E-02	3.46E-05	3.28E-03
RS4806860	1.10E-01	1.55E-01	7.31E-05	3.34E-03
RS2972515	6.72E-02	5.43E-02	7.19E-04	8.58E-03
RS601338	-9.45E-02	-8.53E-02	1.91E-05	9.96E-03

B.15 Chromosome 20

For chromosome 20, from 16936 SNPs, the spike method produced 1171 non-zero variables, lasso method produced 1101 and EBEN method produced 568.

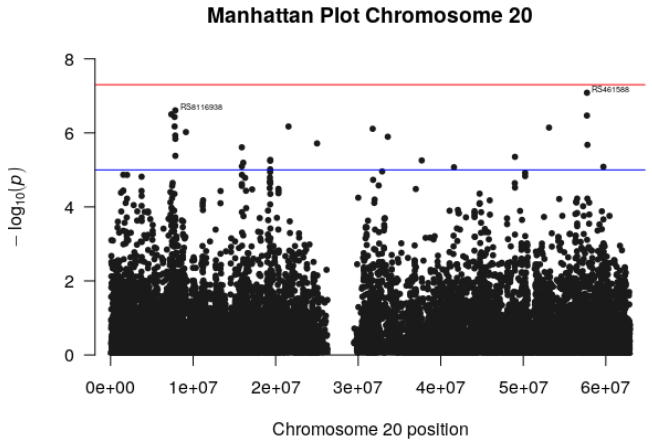

Figure 56: A Manhattan plot with SNPs rs8116938 and rs461588 highlighted

Table 83: Chromosome 20 - Top ten SNPs ranked for each method

Chromosome 20					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS7271722	21712022	1	3	146	0
RS2281496	3235916	2	2	97	0
RS6042797	14562460	3	44	501	0
RS16982896	21699563	4	14	9944	0
RS6023308	53120052	5	1	8	1
RS6033981	14572059	6	23	840	0
RS10485607	46747298	7	8	9088	0
RS8115004	46742514	8	40	6448	0
RS2424407	21745114	9	159	3575	435
RS6028092	59862019	10	4	2268	0
RS212563	54067278	17	5	8196	203
RS6105463	15740824	22	6	13894	396
RS7509151	38841014	29	7	15177	0
RS6103988	43686734	18	9	2652	0
RS17304572	31761919	34	10	12043	0
RS461588	57719291	89	47	1	0
RS8116938	7838533	0	0	2	414
RS6026721	57712488	0	0	3	0
RS11908000	7358423	0	0	4	0
RS6077251	7752366	90	60	5	39
RS2092380	7756027	0	0	6	0
RS6082457	21556715	496	108	7	480
RS7268823	31755539	0	0	9	0
RS6039399	9127494	0	107	10	0
RS2868890	45450419	91	82	372	2
RS6020482	48988571	124	173	660	3
RS235753	6769533	347	400	1575	4
RS2426960	59493906	377	455	315	5
RS10485438	44708638	169	156	58	6
RS6118380	8892600	133	194	305	7
RS326826	56509185	62	266	321	8
RS2747554	32299042	65	303	2205	9
RS4813344	18830812	132	244	2412	10

Table 84: Chromosome 20 - Top ten SNPs for each method

Chromosome 20				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS7271722	-2.98E+00	-7.46E-01	2.60E-04	0
RS2281496	1.70E+00	8.67E-01	1.29E-04	0
RS6042797	1.21E+00	1.72E-01	3.33E-03	0
RS16982896	1.19E+00	3.18E-01	4.73E-01	0
RS6023308	-1.18E+00	-9.24E-01	6.92E-07	8.99E-04
RS6033981	1.16E+00	2.32E-01	8.54E-03	0
RS10485607	1.15E+00	4.57E-01	4.15E-01	0
RS8115004	1.14E+00	1.80E-01	2.45E-01	0
RS2424407	1.11E+00	6.53E-02	9.21E-02	5.97E-01
RS6028092	1.06E+00	5.90E-01	4.43E-02	0
RS212563	8.60E-01	5.42E-01	3.54E-01	3.65E-01
RS6105463	7.97E-01	5.00E-01	7.64E-01	5.50E-01
RS7509151	-7.15E-01	-4.67E-01	8.63E-01	0
RS6103988	8.44E-01	4.55E-01	5.71E-02	0
RS17304572	6.79E-01	4.44E-01	6.32E-01	0
RS461588	1.49E-01	1.67E-01	7.51E-08	0
RS8116938	0	0	2.37E-07	5.78E-01
RS6026721	0	0	2.98E-07	0
RS11908000	0	0	3.04E-07	0
RS6077251	-1.32E-01	-1.36E-01	3.41E-07	1.48E-01
RS2092380	0	0	6.12E-07	0
RS6082457	-2.08E-02	-8.26E-02	6.30E-07	6.54E-01
RS7268823	0	0	8.56E-07	0
RS6039399	0	8.99E-02	9.34E-07	0
RS2868890	1.20E-01	9.98E-02	2.03E-03	8.34E-03
RS6020482	6.50E-02	5.96E-02	5.44E-03	3.79E-02
RS235753	3.10E-02	2.89E-02	2.38E-02	4.83E-02
RS2426960	2.91E-02	2.54E-02	1.45E-03	4.87E-02
RS10485438	5.17E-02	6.62E-02	4.23E-05	5.14E-02
RS6118380	6.29E-02	5.77E-02	1.35E-03	5.99E-02
RS326826	-7.24E-02	-4.35E-02	1.49E-03	6.06E-02
RS2747554	-7.16E-02	-3.93E-02	4.27E-02	6.91E-02
RS4813344	6.33E-02	4.69E-02	4.81E-02	7.30E-02

B.16 Chromosome 21

For chromosome 21, from 9324 SNPs, spike method produced 1124 non-zero variables, lasso method produced 1556 and the EBEN method produced 725.

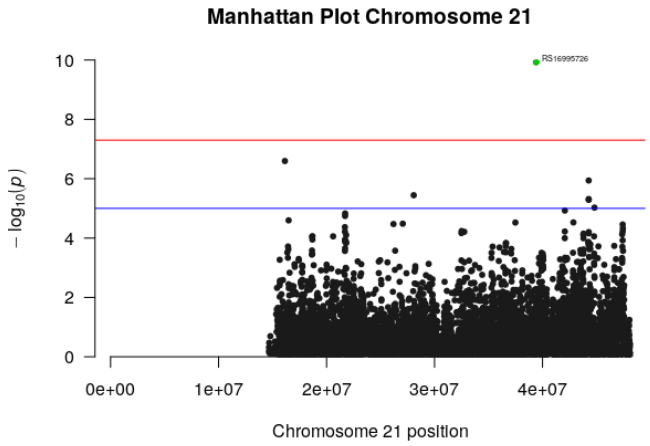

Figure 57: A Manhattan plot with SNP rs16995726 highlighted

Table 85: Chromosome 21 - Top ten SNPs ranked for each method

Chromosome 21					
SNP	Position	SPIKE	LASSO	FREQ	EBEN
RS2828660	25334999	1	9	1682	15
RS587087	44835301	2	48	8544	0
RS476861	44835508	3	69	5292	0
RS8132809	43715801	4	1	198	0
RS2828661	25335327	5	43	2954	0
RS8128762	41907090	6	3	7748	21
RS1997579	16165965	7	58	3154	94
RS2827484	23776582	8	22	3139	14
RS9981655	16186006	9	55	5082	32
RS373477	44453278	10	113	902	7
RS17210142	17815106	11	2	1830	12
RS17000348	26125207	12	4	6689	0
RS762298	23399276	49	5	9243	0
RS11088070	29871279	0	6	2027	0
RS2823279	16807898	39	7	384	0
RS11910928	43716241	16	8	516	0
RS7282856	41891932	29	10	958	0
RS16995726	39400163	43	17	1	0
RS2822918	16141341	55	64	2	1
RS8133752	44271989	61	1001	3	6
RS16978820	28070331	0	853	4	0
RS6586252	44276387	0	0	5	0
RS15736	44273858	0	0	6	0
RS2838295	44808303	364	1067	7	267
RS2837785	42066780	164	361	8	53
RS1487935	21719802	0	0	9	170
RS2826190	21714363	163	1212	10	706
RS2827057	23185634	19	53	707	2
RS9981501	18082155	54	45	8768	3
RS9982633	17419904	60	118	98	4
RS8134569	30164122	33	66	7308	5
RS13052940	23218791	56	60	1375	8
RS2830935	28799489	100	230	7897	9
RS2236674	45880800	142	353	2491	10

Table 86: Chromosome 21 - Top ten SNPs for each method

Chromosome 21				
SNP	SPIKE Coefficient	LASSO Coefficient	FREQ p-value	EBEN p-value
RS2828660	1.36E+00	6.83E-01	7.46E-02	1.33E-03
RS587087	1.27E+00	4.16E-01	8.89E-01	0
RS476861	1.27E+00	3.41E-01	4.57E-01	0
RS8132809	1.22E+00	1.08E+00	1.76E-03	0
RS2828661	1.16E+00	4.52E-01	1.85E-01	0
RS8128762	1.08E+00	9.41E-01	7.79E-01	2.09E-03
RS1997579	-1.05E+00	-3.85E-01	2.05E-01	3.74E-02
RS2827484	1.05E+00	5.64E-01	2.03E-01	1.00E-03
RS9981655	1.04E+00	3.98E-01	4.32E-01	6.33E-03
RS373477	-9.99E-01	-2.39E-01	2.56E-02	2.12E-04
RS17210142	9.82E-01	9.97E-01	8.56E-02	7.26E-04
RS17000348	8.86E-01	9.25E-01	6.42E-01	0
RS762298	5.08E-01	9.01E-01	9.89E-01	0
RS11088070	0	-8.51E-01	1.00E-01	0
RS2823279	6.58E-01	7.48E-01	6.08E-03	0
RS11910928	9.02E-01	6.97E-01	9.71E-03	0
RS7282856	7.39E-01	6.75E-01	2.82E-02	0
RS16995726	-6.29E-01	-6.13E-01	1.01E-10	0
RS2822918	-1.87E-01	-3.58E-01	1.37E-07	6.32E-06
RS8133752	1.18E-01	-5.25E-02	1.66E-06	1.35E-04
RS16978820	0	-6.13E-02	3.98E-06	0
RS6586252	0	0	5.63E-06	0
RS15736	0	0	5.74E-06	0
RS2838295	-3.22E-02	-4.84E-02	9.88E-06	1.55E-01
RS2837785	5.23E-02	1.19E-01	1.52E-05	1.88E-02
RS1487935	0	0	1.52E-05	9.03E-02
RS2826190	-5.15E-02	-4.18E-02	1.82E-05	6.65E-01
RS2827057	8.84E-01	4.02E-01	1.68E-02	7.52E-06
RS9981501	4.00E-01	4.49E-01	9.18E-01	4.47E-05
RS9982633	1.61E-01	2.47E-01	5.48E-04	9.26E-05
RS8134569	7.18E-01	3.47E-01	7.23E-01	1.09E-04
RS13052940	1.64E-01	3.77E-01	5.31E-02	2.78E-04
RS2830935	6.93E-02	1.57E-01	7.99E-01	5.66E-04
RS2236674	-5.67E-02	-1.21E-01	1.40E-01	5.92E-04

C Appendix: Associated SNPs

Appendix C is in tabular form and shows the associated SNPs from Bentham et al study that have been implemented into this study.

C.1 Bentham et al 2015 Associated SNPs featured in this thesis

Table 87 shows the SNPs that were found to be associated with SLE in Bentham et al original study [4].

Table 87: Bentham et al 2015 Associated SNPs

Associated SNPs with closely located gene and calculated p-value			
SNP	Chr	GENE	P-VALUE
rs2476601	1	PTPN22	8.34E-13
rs1801274	1	FCGR2A	6.05E-11
rs704840	1	TNFSF4	1.65E-13
rs17849501	1	SMG7,NCF2	1.63E-59
rs3024505	1	IL10	2.55E-03
rs9782955	1	LYST	5.58E-04
rs6740462	2	SPRED2	2.31E-08
rs2111485	2	IFIH1	3.44E-06
rs11889341	2	STAT4	1.17E-65
rs3768792	2	IKZF2	2.35E-08
rs9311676	3	ABHD6,PXK	5.37E-06
rs564799	3	IL12A	1.15E-06
rs10028805	4	BANK1	4.50E-10
rs7726414	5	TCF7,SKP1	9.17E-10
rs10036748	5	TNIP1	2.83E-18
rs2431697	5	MIR146A	3.23E-14
rs1270942	6	MHC	1.70E-101
rs9462027	6	UHRF1BP1	1.80E-05
rs6568431	6	PRDM1,ATG5	4.33E-12
rs6932056	6	TNFAIP3	1.23E-16
rs849142	7	JAZF1	3.49E-05
rs4917014	7	IKZF1	4.10E-05
rs10488631	7	IRF5	2.66E-44
rs2736340	8	BLK	2.14E-16
rs2663052	10	WDFY4	1.59E-08
rs4948496	10	ARID5B	1.17E-06
rs12802200	11	IRF7	8.43E-09
rs2732549	11	CD44	1.31E-10
rs3794060	11	DHCR7,NADSYN1	1.13E-04
rs7941765	11	ETS1,FLI1	9.82E-07
rs10774625	12	SH2B3	9.47E-08
rs1059312	12	SLC15A4	3.20E-06
rs4902562	14	RAD51B	4.85E-05
rs2289583	15	CSK	9.35E-09
rs11644034	16	IRF8	1.25E-15
rs2941509	17	IKZF3	4.32E-06
rs2304256	19	TYK2	2.34E-12
rs7444	22	UBE2L3	1.30E-13

References

[1] N. H. G. R. Institute, "Chromosome facts: Source - national human genome researchinstitute." https://www.genome.gov/, 2021.

[2] K. B. Hanscombe, D. L. Morris, J. A. Noble, A. T. Dilthey, P. Tombleson, K. M. Kaufman, M. Comeau, C. D. Langefeld, M. E. Alarcon-Riquelme, P. M. Gaffney, *et al.*, "Genetic fine mapping of systemic lupus erythematosus mhc associations in europeans and african americans," *Human molecular genetics*, vol. 27, no. 21, pp. 3813–3824, 2018.

[3] Illumina, "An unbiased view of the entire human genome." https:emea.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing/human.html, 2021.

[4] J. Bentham, D. L. Morris, D. S. C. Graham, C. L. Pinder, P. Tombleson, T. W. Behrens, J. Martín, B. P. Fairfax, J. C. Knight, L. Chen, *et al.*, "Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus," *Nature genetics*, vol. 47, no. 12, pp. 1457–1464, 2015.

[5] T. Strachan, A. P. Read, *et al.*, "Human molecular genetics," *Chromosome Research*, vol. 4, pp. 475–475, 1996.

[6] yourgenome.org, "What is a genetic disorder?." https/www.yourgenome.org/facts/what-is-a-genetic-disorder, 2021.

[7] G. Hom, R. R. Graham, B. Modrek, K. E. Taylor, W. Ortmann, S. Garnier, A. T. Lee, S. A. Chung, R. C. Ferreira, P. K. Pant, *et al.*, "Association of systemic lupus erythematosus with c8orf13–blk and itgam–itgax," *New England Journal of Medicine*, vol. 358, no. 9, pp. 900–909, 2008.

[8] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 499–511, 2010.

[9] O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, and J. Marchini, "Haplotype estimation using sequencing reads," *The American Journal of Human Genetics*, vol. 93, no. 4, pp. 687–696, 2013.

[10] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, p. 56, 2012.

[11] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, "Plink: a tool set for whole-genome association and population-based linkage analyses," *The American journal of human genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[12] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature genetics*, vol. 39, no. 7, pp. 906–913, 2007.

[13] L. Chen, D. L. Morris, and T. J. Vyse, "Genetic advances in systemic lupus erythematosus: an update," *Current opinion in rheumatology*, vol. 29, no. 5, pp. 423–433, 2017.

[14] A. J. Fike, I. Elcheva, and Z. S. Rahman, "The post-gwas era: how to validate the contribution of gene variants in lupus," *Current rheumatology reports*, vol. 21, no. 1, p. 3, 2019.

[15] D. Jiang and M. Wang, "Recent developments in statistical methods for gwas and high-throughput sequencing association studies of complex traits," *Biostatistics & Epidemiology*, vol. 2, no. 1, pp. 132–159, 2018.

[16] F. Weiling, "Historical study: Johann gregor mendel 1822–1884," *American journal of medical genetics*, vol. 40, no. 1, pp. 1–25, 1991.

[17] H. Roest Crollius, O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer, C. Fizames, P. Wincker, P. Brottier, F. Quétier, *et al.*, "Estimate of human gene number provided by genome-wide analysis using tetraodon nigroviridis dna sequence.," *Nature genetics*, vol. 25, no. 2, 2000.

[18] N. H. G. R. Institute, "Chromosome facts: Source - national human genome researchinstitute." https://www.genome.gov/, 2021.

[19] Medlineplus, "What are single nucleotide polymorphisms (snps)?." https://www.medlineplus.gov/genetics/understanding/genomicresearch/snp/, 2021.

[20] K. Ozaki, Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, H. Sato, M. Hori, Y. Nakamura, *et al.*, "Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction," *Nature genetics*, vol. 32, no. 4, pp. 650–654, 2002.

[21] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, *et al.*, "Complement factor h polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[22] M. C. Mills and C. Rahal, "A scientometric review of genome-wide association studies," *Communications biology*, vol. 2, no. 1, pp. 1–11, 2019.

[23] S. H. Jiang, V. Athanasopoulos, J. I. Ellyard, A. Chuah, J. Cappello, A. Cook, S. B. Prabhu, J. Cardenas, J. Gu, M. Stanley, *et al.*, "Functional rare and low frequency variants in blk and bank1 contribute to human lupus," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[24] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[25] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.

[26] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, *et al.*, "The international hapmap project," 2003.

[27] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.

[28] J. J. Chen, "The hardy-weinberg principle and its applications in modern population genetics," *Frontiers in Biology*, vol. 5, no. 4, pp. 348–353, 2010.

[29] P. Armitage, "Tests for linear trends in proportions and frequencies," *Biometrics*, vol. 11, no. 3, pp. 375–386, 1955.

[30] W. G. Cochran, "Some methods for strengthening the common χ 2 tests," *Biometrics*, vol. 10, no. 4, pp. 417–451, 1954.

[31] N. Horita and T. Kaneko, "Genetic model selection for a case–control study and a meta-analysis," *Meta gene*, vol. 5, pp. 1–8, 2015.

[32] L. Bomba, K. Walter, and N. Soranzo, "The impact of rare and low-frequency genetic variants in common disease," *Genome biology*, vol. 18, no. 1, pp. 1–17, 2017.

[33] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, "Common snps explain a large proportion of the heritability for human height," *Nature genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[34] B. Maher, "The case of the missing heritability: when scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. but they were nowhere to be seen. brendan maher shines a light on six places where the missing loot could be stashed away," *Nature*, vol. 456, no. 7218, pp. 18–22, 2008.

[35] R. D. Hernandez, L. H. Uricchio, K. Hartman, C. Ye, A. Dahl, and N. Zaitlen, "Ultrarare variants drive substantial cis heritability of human gene expression," *Nature genetics*, vol. 51, no. 9, pp. 1349–1355, 2019.

[36] lupus.org, "Lupus symptoms." https://www.lupus.org/resources/common-symptoms-of-lupus, 2021.

[37] L. F. of America, "What is lupus?." https://www.lupus.org/resources/what-is-lupus, 2021.

[38] B. H. Hahn and D. J. Wallace, *Dubois' Lupus Erythematosus and Related Syndromes*. Elsevier/Saunders, 2013.

[39] G. Roper, "Lupus awareness survey for the lupus foundation of america [executive summary report]," *Washington, DC. GfK Roper Public Affairs & Corporate Communications*, 2012.

[40] G. J. Pons-Estel, G. S. Alarcón, L. Scofield, L. Reinlib, and G. S. Cooper, "Understanding the epidemiology and progression of systemic lupus erythematosus," in *Seminars in arthritis and rheumatism*, vol. 39, pp. 257–268, Elsevier, 2010.

[41] J. B. Harley, M. E. Alarcón-Riquelme, L. A. Criswell, C. O. Jacob, R. P. Kimberly, K. L. Moser, B. P. Tsao, T. J. Vyse, and C. D. Langefeld, "Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in itgam, pxk, kiaa1542 and other loci," *Nature genetics*, vol. 40, no. 2, pp. 204–210, 2008.

[42] rheumatology.org, "Lupus." https/www.rheumatology.org/I-Am-A/Patient-Caregiver/Diseases-Conditions/Lupus, 2021.

[43] D. L. Morris, Y. Sheng, Y. Zhang, Y.-F. Wang, Z. Zhu, P. Tombleson, L. Chen, D. S. C. Graham, J. Bentham, A. L. Roberts, *et al.*, "Genome-wide association meta-analysis in chinese and european individuals identifies ten new loci associated with systemic lupus erythematosus," *Nature genetics*, vol. 48, no. 8, p. 940, 2016.

[44] M. S. Consortium *et al.*, "Complete sequence and gene map of a human major histocompatibility complex," *Natur*, vol. 401, no. 6756, pp. 921–923, 1999.

[45] F. C. Grumet, A. Coukell, J. G. Bodmer, W. F. Bodmer, and H. O. McDevitt, "Histocompatibility (hl-a) antigens associated with systemic lupus erythematosus: a possible genetic predisposition to disease," *New England Journal of Medicine*, vol. 285, no. 4, pp. 193–196, 1971.

[46] H. Waters, P. Konrad, and R. L. Walford, "The distribution of hl-a histocompatibility factors and genes in patients with systemic lupus erythematosus," *Tissue Antigens*, vol. 1, no. 2, pp. 68–73, 1971.

[47] D. Morris, M. Fernando, K. Taylor, S. Chung, J. Nititham, M. Alarcon-Riquelme, L. Barcellos, T. Behrens, C. Cotsapas, P. Gaffney, *et al.*, "Mhc associations with clinical and autoantibody manifestations in european sle," *Genes & Immunity*, vol. 15, no. 4, pp. 210–217, 2014.

[48] M. M. Fernando, C. R. Stevens, P. C. Sabeti, E. C. Walsh, A. J. McWhinnie, A. Shah, T. Green, J. D. Rioux, and T. J. Vyse, "Identification of two independent risk factors for lupus within the mhc in united kingdom families," *PLoS Genet*, vol. 3, no. 11, p. e192, 2007.

[49] C. Vandiedonck and J. C. Knight, "The human major histocompatibility complex as a paradigm in genomics research," *Briefings in Functional Genomics and Proteomics*, vol. 8, no. 5, pp. 379–394, 2009.

[50] N. Bottini, T. Vang, F. Cucca, and T. Mustelin, "Role of ptpn22 in type 1 diabetes and other autoimmune diseases," in *Seminars in immunology*, vol. 18, pp. 207–213, Elsevier, 2006.

[51] K. Su, J. Wu, J. C. Edberg, X. Li, P. Ferguson, G. S. Cooper, C. D. Langefeld, and R. P. Kimberly, "A promoter haplotype of the immunoreceptor tyrosine-based inhibitory motif-bearing fcγriib alters receptor expression and associates with autoimmunity. i. regulatory fcgr2b polymorphisms and their association with systemic lupus erythematosus," *The Journal of Immunology*, vol. 172, no. 11, pp. 7186–7191, 2004.

[52] E. F. Remmers, R. M. Plenge, A. T. Lee, R. R. Graham, G. Hom, T. W. Behrens, P. I. De Bakker, J. M. Le, H.-S. Lee, F. Batliwalla, *et al.*, "Stat4 and the risk of rheumatoid arthritis and systemic lupus erythematosus," *New England Journal of Medicine*, vol. 357, no. 10, pp. 977–986, 2007.

[53] L. Prokunina, C. Castillejo-López, F. Öberg, I. Gunnarsson, L. Berg, V. Magnusson, A. J. Brookes, D. Tentler, H. Kristjansdóttir, G. Gröndal, *et al.*, "A regulatory polymorphism in pdcd1 is associated with susceptibility to systemic lupus erythematosus in humans," *Nature genetics*, vol. 32, no. 4, pp. 666–669, 2002.

[54] J.-W. Hur, Y.-K. Sung, H. D. Shin, B. L. Park, H. S. Cheong, and S.-C. Bae, "Trex1 polymorphisms associated with autoantibodies in patients with systemic lupus erythematosus," *Rheumatology international*, vol. 28, no. 8, pp. 783–789, 2008.

[55] C. Wong, L. Lit, L. Tam, E. Li, and C. Lam, "Elevation of plasma osteopontin concentration is correlated with disease activity in patients with systemic lupus erythematosus," *Rheumatology*, vol. 44, no. 5, pp. 602–606, 2005.

[56] S. V. Kozyrev, A.-K. Abelson, J. Wojcik, A. Zaghlool, M. P. L. Reddy, E. Sanchez, I. Gunnarsson, E. Svenungsson, G. Sturfelt, A. Jönsen, *et al.*, "Functional variants in the b-cell gene bank1 are associated with systemic lupus erythematosus," *Nature genetics*, vol. 40, no. 2, pp. 211–216, 2008.

[57] P. Brennan, A. Hajeer, K. R. Ong, J. Worthington, S. John, W. Thomson, A. Silman, and B. Ollier, "Allelic markers close to prolactin are associated with hla-drb1 susceptibility alleles among women with rheumatoid arthritis and systemic lupus erythematosus," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 40, no. 8, pp. 1383–1386, 1997.

[58] R. R. Graham, S. V. Kozyrev, E. C. Baechler, M. P. L. Reddy, R. M. Plenge, J. W. Bauer, W. A. Ortmann, T. Koeuth, M. F. G. Escribano, B. Pons-Estel, *et al.*, "A common haplotype of interferon regulatory factor 5 (irf5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus," *Nature genetics*, vol. 38, no. 5, pp. 550–555, 2006.

[59] R. R. Graham, C. Cotsapas, L. Davies, R. Hackett, C. J. Lessard, J. M. Leon, N. P. Burtt, C. Guiducci, M. Parkin, C. Gates, *et al.*, "Genetic variants near tnfaip3 on 6q23 are associated with systemic lupus erythematosus," *Nature genetics*, vol. 40, no. 9, pp. 1059–1061, 2008.

[60] V. Gateva, J. K. Sandling, G. Hom, K. E. Taylor, S. A. Chung, X. Sun, W. Ortmann, R. Kosoy, R. C. Ferreira, G. Nordmark, *et al.*, "A large-scale replication study identifies tnip1, prdm1, jazf1, uhrf1bp1 and il10 as risk loci for systemic lupus erythematosus," *Nature genetics*, vol. 41, no. 11, pp. 1228–1233, 2009.

[61] S. A. Chung, K. E. Taylor, R. R. Graham, J. Nititham, A. T. Lee, W. A. Ortmann, C. O. Jacob, M. E. Alarcón-Riquelme, B. P. Tsao, J. B. Harley, *et al.*, "Differential genetic associations for systemic lupus erythematosus based on anti–dsdna autoantibody production," *PLoS Genet*, vol. 7, no. 3, p. e1001323, 2011.

[62] D. L. Armstrong, R. Zidovetzki, M. E. Alarcón-Riquelme, B. P. Tsao, L. A. Criswell, R. P. Kimberly, J. B. Harley, K. L. Sivils, T. J. Vyse, P. M. Gaffney, *et al.*, "Gwas identifies novel sle susceptibility genes and explains the association of the hla region," *Genes & Immunity*, vol. 15, no. 6, pp. 347–354, 2014.

[63] C. D. Langefeld, H. C. Ainsworth, D. S. C. Graham, J. A. Kelly, M. E. Comeau, M. C. Marion, T. D. Howard, P. S. Ramos, J. A. Croker, D. L. Morris, *et al.*, "Transancestral mapping and genetic load in systemic lupus erythematosus," *Nature communications*, vol. 8, no. 1, pp. 1–18, 2017.

[64] A. Julià, F. J. López-Longo, J. J. P. Venegas, S. Bonàs-Guarch, À. Olivé, J. L. Andreu, M. Á. Aguirre-Zamorano, P. Vela, J. M. Nolla, J. L. M. de la Fuente, *et al.*, "Genome-wide association study meta-analysis identifies five new loci for systemic lupus erythematosus," *Arthritis research & therapy*, vol. 20, no. 1, pp. 1–10, 2018.

[65] H. Zhang, Y. Zhang, Y.-F. Wang, D. Morris, N. Hirankarn, Y. Sheng, J. Shen, H.-F. Pan, J. Yang, S. Yang, *et al.*, "Meta-analysis of gwas on both chinese and european populations identifies gpr173 as a novel x chromosome susceptibility gene for sle," *Arthritis research & therapy*, vol. 20, no. 1, pp. 1–8, 2018.

[66] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[67] G. Rupert Jr *et al.*, *Simultaneous statistical inference*. Springer Science & Business Media, 2012.

[68] T. V. Perneger, "What's wrong with bonferroni adjustments," *Bmj*, vol. 316, no. 7139, pp. 1236–1238, 1998.

[69] S. Nakagawa, "A farewell to bonferroni: the problems of low statistical power and publication bias," *Behavioral ecology*, vol. 15, no. 6, pp. 1044–1045, 2004.

[70] D. Heath and W. Sudderth, "De finetti's theorem on exchangeable variables," *The American Statistician*, vol. 30, no. 4, pp. 188–189, 1976.

[71] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.

[72] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," 1970.

[73] T. P. Minka, "Expectation propagation for approximate bayesian inference," *arXiv preprint arXiv:1301.2294*, 2013.

[74] J. Paisley, D. Blei, and M. Jordan, "Variational bayesian inference with stochastic search," *arXiv preprint arXiv:1206.6430*, 2012.

[75] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[76] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[77] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[78] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[79] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[80] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[81] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[82] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[83] V. Ročková and E. I. George, "The spike-and-slab lasso," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 431–444, 2018.

[84] A. Huang, S. Xu, and X. Cai, "Empirical bayesian elastic net for multiple quantitative trait locus mapping," *Heredity*, vol. 114, no. 1, pp. 107–115, 2015.

[85] O. Kohannim, D. P. Hibar, J. L. Stein, N. Jahanshad, X. Hua, P. Rajagopalan, A. Toga, C. R. Jack Jr, M. W. Weiner, G. I. De Zubicaray, *et al.*, "Discovery and replication of gene influences on brain structure using lasso regression," *Frontiers in neuroscience*, vol. 6, p. 115, 2012.

[86] Z. Tang, Y. Shen, X. Zhang, and N. Yi, "The spike-and-slab lasso generalized linear models for prediction and associated genes detection," *Genetics*, vol. 205, no. 1, pp. 77–88, 2017.

[87] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[88] X. Lu, E. E. Zoller, M. T. Weirauch, Z. Wu, B. Namjou, A. H. Williams, J. T. Ziegler, M. E. Comeau, M. C. Marion, S. B. Glenn, *et al.*, "Lupus risk variant increases pstat1 binding and decreases ets1 expression," *The American Journal of Human Genetics*, vol. 96, no. 5, pp. 731–739, 2015.

[89] B. M. Neale and S. Purcell, "The positives, protocols, and perils of genome-wide association," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 147, no. 7, pp. 1288–1294, 2008.

[90] T. Hastie and J. Qian, "Glmnet vignette," *Retrieved June*, vol. 9, no. 2016, pp. 1–30, 2014.

[91] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

[92] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[93] F. B. Lempers and A. S. Louter, "An extension of the table of the student distribution," *Journal of the American Statistical Association*, vol. 66, no. 335, pp. 503–503, 1971.

[94] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the american statistical association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[95] E. I. George and R. E. McCulloch, "Variable selection via gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[96] X. Cai, A. Huang, and S. Xu, "Fast empirical bayesian lasso for multiple quantitative trait locus mapping," *BMC bioinformatics*, vol. 12, no. 1, p. 211, 2011.

[97] A. Huang, S. Xu, and X. Cai, "Empirical bayesian lasso-logistic regression for multiple binary trait locus mapping," *BMC genetics*, vol. 14, no. 1, p. 5, 2013.

[98] G. Orozco, E. Sánchez, M. A. González-Gay, M. A. López-Nevot, B. Torres, R. Cáliz, N. Ortego-Centeno, J. Jiménez-Alonso, D. Pascual-Salcedo, A. Balsa, *et al.*, "Association of a functional single-nucleotide polymorphism of ptpn22, encoding lymphoid protein phosphatase, with rheumatoid arthritis and systemic lupus erythematosus," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 52, no. 1, pp. 219–224, 2005.

[99] C. Kyogoku, H. M. Dijstelbloem, N. Tsuchiya, Y. Hatta, H. Kato, A. Yamaguchi, T. Fukazawa, M. D. Jansen, H. Hashimoto, J. G. van de Winkel, *et al.*, "Fcγ receptor gene polymorphisms in japanese patients with systemic lupus erythematosus: contribution of fcgr2b to genetic susceptibility," *Arthritis & Rheumatism*, vol. 46, no. 5, pp. 1242–1254, 2002.

[100] J.-E. Martin, S. Assassi, L.-M. Diaz-Gallo, J. C. Broen, C. P. Simeon, I. Castellvi, E. Vicente-Rabaneda, V. Fonollosa, N. Ortego-Centeno, M. A. González-Gay, *et al.*, "A systemic sclerosis and systemic lupus erythematosus pan-metagwas reveals new shared susceptibility loci," *Human molecular genetics*, vol. 22, no. 19, pp. 4021–4029, 2013.

[101] Y. Wang, S. D. Bos, H. F. Harbo, W. K. Thompson, A. J. Schork, F. Bettella, A. Witoelar, B. A. Lie, W. Li, L. K. McEvoy, *et al.*, "Genetic overlap between multiple sclerosis and several cardiovascular disease risk factors," *Multiple Sclerosis Journal*, vol. 22, no. 14, pp. 1783–1793, 2016.

[102] J.-W. Han, H.-F. Zheng, Y. Cui, L.-D. Sun, D.-Q. Ye, Z. Hu, J.-H. Xu, Z.-M. Cai, W. Huang, G.-P. Zhao, *et al.*, "Genome-wide association study in a chinese han population identifies nine new susceptibility loci for systemic lupus erythematosus," *Nature genetics*, vol. 41, no. 11, pp. 1234–1237, 2009.

[103] Y.-j. Sheng, J.-h. Xu, Y.-g. Wu, X.-b. Zuo, J.-p. Gao, Y. Lin, Z.-w. Zhu, L.-l. Wen, C. Yang, L. Liu, *et al.*, "Association analyses confirm five susceptibility loci for systemic lupus erythematosus in the han chinese population," *Arthritis research & therapy*, vol. 17, no. 1, pp. 1–7, 2015.

[104] J. D. Cooper, M. J. Simmonds, N. M. Walker, O. Burren, O. J. Brand, H. Guo, C. Wallace, H. Stevens, G. Coleman, W. T. C. C. Consortium, *et al.*, "Seven newly identified loci for autoimmune thyroid disease," *Human molecular genetics*, vol. 21, no. 23, pp. 5202–5208, 2012.

[105] G. Orozco, S. Viatte, J. Bowes, P. Martin, A. G. Wilson, A. W. Morgan, S. Steer, P. Wordsworth, L. J. Hocking, U. R. A. G. Consortium, *et al.*, "Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended uk genome-wide association study," *Arthritis & Rheumatology*, vol. 66, no. 1, pp. 24–30, 2014.

[106] W. Yang, P. Ng, M. Zhao, N. Hirankarn, C. Lau, C. Mok, T. Chan, R. Wong, K. Lee, M. Mok, *et al.*, "Population differences in sle susceptibility genes: Stat4 and blk, but not pxk, are associated with systemic lupus erythematosus in hong kong chinese," *Genes & Immunity*, vol. 10, no. 3, pp. 219–226, 2009.

[107] A. Márquez, L. Vidal-Bralo, L. Rodríguez-Rodríguez, M. A. González-Gay, A. Balsa, I. González-Álvaro, P. Carreira, N. Ortego-Centeno, M. M. Ayala-Gutiérrez, F. J. García-Hernández, *et al.*, "A combined large-scale meta-analysis identifies cog6 as a novel shared risk locus for rheumatoid arthritis and systemic lupus erythematosus," *Annals of the rheumatic diseases*, vol. 76, no. 1, pp. 286–294, 2017.

[108] M. E. Alarcón-Riquelme, J. T. Ziegler, J. Molineros, T. D. Howard, A. Moreno-Estrada, E. Sánchez-Rodríguez, H. C. Ainsworth, P. Ortiz-Tello, M. E. Comeau, A. Rasmussen, *et al.*, "Genome-wide association study in an amerindian ancestry population reveals novel systemic lupus erythematosus risk loci and the role of european admixture," *Arthritis & rheumatology*, vol. 68, no. 4, pp. 932–943, 2016.

[109] X. Zuo, L. Sun, X. Yin, J. Gao, Y. Sheng, J. Xu, J. Zhang, C. He, Y. Qiu, G. Wen, *et al.*, "Whole-exome snp array identifies 15 new susceptibility loci for psoriasis," *Nature communications*, vol. 6, no. 1, pp. 1–7, 2015.

[110] R.-P. Dong, A. Kimura, H. Hashimoto, M. Akizuki, Y. Nishimura, and T. Sasazuki, "Difference in hla-linked genetic background between mixed connective tissue disease and systemic lupus erythematosus," *Tissue antigens*, vol. 41, no. 1, pp. 20–25, 1993.

[111] J. D. Reveille, "Genetic studies in the rheumatic diseases: present status and implications for the future.," *The Journal of Rheumatology Supplement*, vol. 72, pp. 10–13, 2005.

[112] H.-C. Yang, L.-C. Chang, Y.-J. Liang, C.-H. Lin, and P.-L. Wang, "A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human major histocompatibility complex," *PloS one*, vol. 7, no. 4, p. e34840, 2012.

[113] C. Sun, J. E. Molineros, L. L. Looger, X.-j. Zhou, K. Kim, Y. Okada, J. Ma, Y.-y. Qi, X. Kim-Howard, P. Motghare, *et al.*, "High-density genotyping of immune-related loci identifies new sle risk variants in individuals with asian ancestry," *Nature genetics*, vol. 48, no. 3, pp. 323–330, 2016.

[114] C. J. Lessard, S. Sajuthi, J. Zhao, K. Kim, J. A. Ice, H. Li, H. Ainsworth, A. Rasmussen, J. A. Kelly, M. Marion, *et al.*, "Identification of a systemic lupus erythematosus risk locus spanning atg16l2, fchsd2, and p2ry2 in koreans," *Arthritis & Rheumatology*, vol. 68, no. 5, pp. 1197–1209, 2016.

[115] S. K. Nath, B. Namjou, D. Hutchings, C. P. Garriott, C. Pongratz, J. Guthridge, and J. J. James, "Systemic lupus erythematosus (sle) and chromosome 16: confirmation of linkage to 16q12–13 and evidence for genetic heterogeneity," *European journal of human genetics*, vol. 12, no. 8, pp. 668–672, 2004.

[116] H. Lv, M. Zhang, Z. Shang, J. Li, S. Zhang, D. Lian, and R. Zhang, "Genome-wide haplotype association study identify the fgfr2 gene as a risk gene for acute myeloid leukemia," *Oncotarget*, vol. 8, no. 5, p. 7891, 2017.

[117] K. Kiryluk, Y. Li, F. Scolari, S. Sanna-Cherchi, M. Choi, M. Verbitsky, D. Fasel, S. Lata, S. Prakash, S. Shapiro, *et al.*, "Discovery of new risk loci for iga nephropathy implicates genes involved in immunity against intestinal pathogens," *Nature genetics*, vol. 46, no. 11, pp. 1187–1196, 2014.

[118] W. Yang, H. Tang, Y. Zhang, X. Tang, J. Zhang, L. Sun, J. Yang, Y. Cui, L. Zhang, N. Hirankarn, *et al.*, "Meta-analysis followed by replication identifies loci in or near cdkn1b, tet3, cd80, dram1, and arid5b as associated with systemic lupus erythematosus in asians," *The American Journal of Human Genetics*, vol. 92, no. 1, pp. 41–51, 2013.

[119] R. Hou, S. A. Cole, M. Graff, K. Haack, S. Laston, A. G. Comuzzie, N. R. Mehta, K. Ryan, D. L. Cousminer, B. S. Zemel, *et al.*, "Genetic variants affecting bone mineral density and bone mineral content at multiple skeletal sites in hispanic children," *Bone*, vol. 132, p. 115175, 2020.

[120] S. Tuteja, L. Qu, M. Vujkovic, R. L. Dunbar, J. Chen, S. DerOhannessian, and D. J. Rader, "Genetic variants associated with plasma lipids are associated with the lipid response to niacin," *Journal of the American Heart Association*, vol. 7, no. 19, p. e03488, 2018.

[121] S. Wang, I. Adrianto, G. B. Wiley, C. J. Lessard, J. A. Kelly, A. J. Adler, S. B. Glenn, A. H. Williams, J. T. Ziegler, M. E. Comeau, *et al.*, "A functional haplotype of ube2l3 confers risk for systemic lupus erythematosus," *Genes & Immunity*, vol. 13, no. 5, pp. 380–387, 2012.

[122] L. C. Pilling, J. L. Atkins, M. O. Duff, R. N. Beaumont, S. E. Jones, J. Tyrrell, C.-L. Kuo, K. S. Ruth, M. A. Tuke, H. Yaghootkar, *et al.*, "Red blood cell distribution width: genetic evidence for aging pathways in 116,666 volunteers," *PLoS One*, vol. 12, no. 9, p. e0185083, 2017.

[123] C. N. Spracklen, P. Chen, Y. J. Kim, X. Wang, H. Cai, S. Li, J. Long, Y. Wu, Y. X. Wang, F. Takeuchi, *et al.*, "Association analyses of east asian individuals

and trans-ancestry analyses with european individuals reveal new loci associated with cholesterol and triglyceride levels," *Human molecular genetics*, vol. 26, no. 9, pp. 1770–1784, 2017.

[124] N. Yi, Z. Tang, X. Zhang, and B. Guo, "Bhglm: Bayesian hierarchical glms and survival models, with applications to genomics and epidemiology," *Bioinformatics*, vol. 35, no. 8, pp. 1419–1421, 2019.

[125] J. E. Molineros, W. Yang, X.-j. Zhou, C. Sun, Y. Okada, H. Zhang, K. Heng Chua, Y.-L. Lau, Y. Kochi, A. Suzuki, *et al.*, "Confirmation of five novel susceptibility loci for systemic lupus erythematosus (sle) and integrated network analysis of 82 sle susceptibility loci," *Human molecular genetics*, vol. 26, no. 6, pp. 1205–1216, 2017.

[126] S. I. Berndt, N. J. Camp, C. F. Skibola, J. Vijai, Z. Wang, J. Gu, A. Nieters, R. S. Kelly, K. E. Smedby, A. Monnereau, *et al.*, "Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia," *Nature communications*, vol. 7, no. 1, pp. 1–9, 2016.

[127] S. Sigurdsson, G. Nordmark, H. H. Göring, K. Lindroos, A.-C. Wiman, G. Sturfelt, A. Jönsen, S. Rantapää-Dahlqvist, B. Möller, J. Kere, *et al.*, "Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus," *The American Journal of Human Genetics*, vol. 76, no. 3, pp. 528–537, 2005.

[128] A. Kawasaki, C. Kyogoku, J. Ohashi, R. Miyashita, K. Hikami, M. Kusaoi, K. Tokunaga, Y. Takasaki, H. Hashimoto, T. W. Behrens, *et al.*, "Association of irf5 polymorphisms with systemic lupus erythematosus in a japanese population: support for a crucial role of intron 1 polymorphisms," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 58, no. 3, pp. 826–834, 2008.

[129] J. Bowes, P. Ho, E. Flynn, F. Ali, H. Marzo-Ortega, L. C. Coates, R. B. Warren, R. McManus, A. W. Ryan, D. Kane, *et al.*, "Comprehensive assessment of rheumatoid arthritis susceptibility loci in a large psoriatic arthritis cohort," *Annals of the rheumatic diseases*, vol. 71, no. 8, pp. 1350–1354, 2012.

[130] P. Gourh, S. K. Agarwal, E. Martin, D. Divecha, B. Rueda, H. Bunting, S. Assassi, G. Paz, S. Shete, T. McNearney, *et al.*, "Association of the c8orf13-blk region with systemic sclerosis in north-american and european populations," *Journal of autoimmunity*, vol. 34, no. 2, pp. 155–162, 2010.

[131] L. Din, M. Sheikh, N. Kosaraju, K. E. Smedby, S. Bernatsky, S. I. Berndt, C. F. Skibola, A. Nieters, S. Wang, J. D. McKay, *et al.*, "Genetic overlap between autoimmune diseases and non-hodgkin lymphoma subtypes," *Genetic epidemiology*, vol. 43, no. 7, pp. 844–863, 2019.

[132] C. J. Lessard, I. Adrianto, J. A. Kelly, K. M. Kaufman, K. M. Grundahl, A. Adler, A. H. Williams, C. J. Gallant, J.-M. Anaya, S.-C. Bae, *et al.*, "Identification of a systemic lupus erythematosus susceptibility locus at 11p13 between pdhx and cd44 in a multiethnic study," *The American Journal of Human Genetics*, vol. 88, no. 1, pp. 83–91, 2011.

[133] L. Wen, C. Zhu, Z. Zhu, C. Yang, X. Zheng, L. Liu, X. Zuo, Y. Sheng, H. Tang, B. Liang, *et al.*, "Exome-wide association study identifies four novel loci for systemic lupus erythematosus in han chinese population," *Annals of the rheumatic diseases*, vol. 77, no. 3, pp. 417–417, 2018.

[134] A. Hellquist, T. M. Järvinen, S. Koskenmies, M. Zucchelli, C. Orsmark-Pietras, L. Berglind, J. Panelius, T. Hasan, H. Julkunen, M. D'Amato, *et al.*, "Evidence for genetic association and interaction between the tyk2 and irf5 genes in systemic lupus erythematosus," *The Journal of Rheumatology*, vol. 36, no. 8, pp. 1631–1638, 2009.

[135] C. Kyogoku, A. Morinobu, K. Nishimura, D. Sugiyama, H. Hashimoto, Y. Tokano, T. Mimori, C. Terao, F. Matsuda, T. Kuno, *et al.*, "Lack of association between tyrosine kinase 2 (tyk2) gene polymorphisms and susceptibility to sle in a japanese population," *Modern rheumatology*, vol. 19, no. 4, pp. 401–406, 2009.

[136] K. Kim, E. E. Brown, C.-B. Choi, M. E. Alarcón-Riquelme, J. A. Kelly, S. B. Glenn, J. O. Ojwang, A. Adler, H.-S. Lee, S. A. Boackle, *et al.*, "Variation in the icam1–icam4–icam5 locus is associated with systemic lupus erythematosus susceptibility in multiple ancestries," *Annals of the rheumatic diseases*, vol. 71, no. 11, pp. 1809–1814, 2012.